Discover Computing

Research

Leveraging large language model as news sentiment predictor in stock markets: a knowledge-enhanced strategy

Weisi Chen[1] · **Wulong Liu**[2] · **Jiaxin Zheng**[2] · **Xu Zhang**[1]

Received: 26 June 2024 / Accepted: 24 April 2025
Published online: 09 May 2025

Abstract
In the fast-evolving artificial intelligence era, the intersection of natural language processing and financial analysis has attracted significant attention, primarily due to its potential to provide valuable insights into financial market behavior. Sentiment analysis of financial news articles is a crucial aspect of this intersection, providing cues about market sentiment that may affect stock price dynamics. Traditional sentiment analysis methods often rely on rules or machine learning algorithms trained on labeled datasets, but these methods face challenges in capturing the context within the text. This paper proposes a framework that incorporates prompt engineering strategies, including a novel Domain Knowledge Chain-of-Thought (DK-CoT) strategy, integrating domain-specific financial knowledge with chain-of-thought reasoning, designed to leverage and enhance the performance of large language models (LLMs) in financial news sentiment analysis. DK-CoT has been compared with various prompt engineering techniques, including zero-shot, few-shot, and chain-of-thought, as well as other benchmark models like BERT and RoBERTa. Through comprehensive experiments and evaluations, we introduce the weighted F1 score as a more practical metric, emphasizing the disproportionate impact of negative news on financial markets, which better reflects real-world financial dynamics, as negative sentiments often lead to more significant market reactions than positive or neutral sentiments. Experimental results have shown that DK-CoT adopted in an LLM called GLM is effective in improving the performance and reliability of financial news sentiment analysis. Our findings provide insights into optimal prompt designs and highlight the importance of incorporating financial knowledge to uplift LLM performance while reducing the need for extensive computational resources and fine-tuning.

Keywords Sentiment analysis · Financial news · Natural language processing · Accessible machine learning · Large language model · Prompt engineering

1 Introduction

1.1 Background

In recent years, the intersection of natural language processing (NLP) and financial analysis has garnered significant attention due to its potential to provide valuable insights into stock market behavior. One crucial aspect of this

✉ Weisi Chen, chenweisi@xmut.edu.cn; Wulong Liu, 2222031204@stu.xmut.edu.cn; Jiaxin Zheng, 2322071052@s.xmut.edu.cn; Xu Zhang, zhangxu@xmut.edu.cn | [1]School of Software Engineering, Xiamen University of Technology, 600 Ligong Road, Houxi County, Jimei District, Xiamen, Fujian, China. [2]School of Computer and Information Engineering, Xiamen University of Technology, 600 Ligong Road, Houxi County, Jimei District, Xiamen, Fujian, China.

intersection is sentiment analysis, which involves extracting and interpreting sentiment or emotion from textual data [1]. Sentiment analysis applied to financial news articles can offer valuable cues about market sentiment, which, in turn, can influence stock price movements.

Traditional approaches to sentiment analysis often rely on machine learning algorithms trained on labeled datasets. However, these methods usually face challenges in capturing the context of financial news articles. LLMs that are pre-trained using large amounts of textual data have proven to be effective in all NLP tasks including sentiment analysis. Prompt engineering techniques offer a promising alternative by providing a structured approach to guide the generation of contextual prompts for LLMs, thereby enhancing their ability to understand and interpret specific text types. The use of prompt engineering in sentiment analysis has shown promise in various domains, including social media and product review analysis. Extending these techniques to financial news sentiment analysis presents an exciting opportunity to improve the accuracy and reliability of results, which can provide better insights into stock price movements [2]. Figure 1 describes how LLMs and prompt engineering could facilitate financial sentiment analysis, and how it can be integrated into downstream modules to analyze stock price movements.

To the best of our knowledge, there is a lack of evaluation on prompt engineering strategies' effectiveness on financial news sentiment analysis. This paper bridges the gap by proposing a novel prompt engineering framework called Domain Knowledge Chain-of-Thought (DK-CoT) and evaluating and comparing prevalent strategies in the context of stock-related news sentiment analysis. By systematically analyzing the performance of various prompt designs and methodologies, we seek to provide insights into the optimal strategies for leveraging prompt engineering to enhance the accuracy of news sentiment analysis.

1.2 Research questions and contributions

The primary objective of this study is to advance the field of financial news sentiment analysis through the development and evaluation of a novel prompt engineering strategy, DK-CoT, versus other prevalent strategies. This paper attempts to address the following research questions:

RQ 1. How can LLMs be effectively leveraged as financial news sentiment predictors to enhance the accuracy and timeliness of stock market predictions?

This paper investigates the potential of LLMs, taking GLM [3] as an example, in financial news sentiment analysis, capturing market sentiments that could influence stock prices effectively. This study aims to explore how these advanced models can be adapted via techniques like prompt engineering and domain knowledge insertion to provide precise and timely sentiment predictions. The findings will also inspire how LLMs can be integrated into existing financial analysis frameworks to provide more accurate and meaningful insights, thereby supporting more informed investment strategies.

RQ 2. How effective is the proposed novel prompt engineering technique in enhancing the accuracy and reliability of sentiment analysis for financial news, compared with other techniques like zero-shot, few-shot, and chain-of-thought (CoT)?

By comparing the performance of techniques such as zero-shot, few-shot, CoT, and the novel DK-CoT strategy, this paper seeks to identify the most effective approach that can enhance the precision and reliability of sentiment analysis results. The findings will provide insights into how structured prompt template designs can guide LLMs to better interpret and predict market sentiments, ultimately contributing to more accurate stock market predictions.

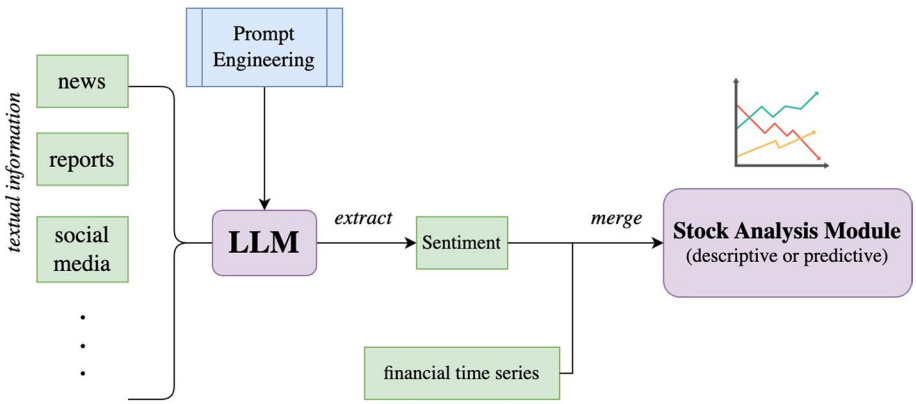


Fig. 1 LLM as financial sentiment predictor

This study makes the following contributions to the financial news sentiment analysis field:

- Introduction and validation of the novel DK-CoT prompt engineering strategy: the paper introduces a novel prompt engineering strategy called Domain Knowledge Chain-of-Thought (DK-CoT), which integrates domain-specific financial knowledge with CoT reasoning. Through comprehensive evaluation, the study demonstrates whether DK-CoT can enhance the accuracy and reliability of sentiment analysis for financial news, outperforming other benchmark models based on BERT and RoBERTa, and other prompting strategies.
- Insights into optimal prompt designs and integration of domain knowledge: the paper provides valuable insights into the optimal strategies for leveraging prompt engineering in sentiment analysis. By systematically comparing various prompt designs and methodologies, the study highlights the importance of incorporating domain-specific financial knowledge to improve LLMs' performance in predicting stock market movements based on financial news sentiment.
- New evaluation metric: the paper introduces the weighted F1 score as a more practical evaluation metric for financial news sentiment analysis. This metric considers the disproportionate impact of negative news on financial markets, aligning the evaluation process with real-world financial dynamics. By assigning higher weights to negative sentiments than to neutral and positive ones, the weighted F1 score provides a different perspective with a more practical assessment of sentiment analysis models' performance.
- Promotion of practical and sustainable artificial intelligence (AI) applications in finance: the findings emphasize the practical applications of advanced prompt engineering techniques in financial decision-making. Additionally, the research promotes sustainable AI practices by showcasing efficient and accessible prompt engineering techniques that reduce the need for extensive computational resources and fine-tuning, making advanced AI capabilities more accessible to a broader audience.

The rest of the paper is structured as follows. Section 2 discusses related work on financial news sentiment analysis, highlighting the evolution of paradigms and techniques. Section 3 illustrates in detail the proposed DK-CoT prompting strategy, along with relevant techniques for achieving it. Section 4 explains the design of the experiment and its setup. Section 5 demonstrates and discusses the experimental results. Finally, Sect. 6 concludes the paper and highlights future research directions.

2 Related work

In the fast-evolving AI era, news sentiment analysis in the financial domain has evolved from rule-based approaches to sophisticated deep learning models.

Early approaches to sentiment analysis primarily relied on rule-based AI. These methods utilize predefined sets of rules and lexicons to identify sentiment-bearing words and phrases within the text. A well-known example is the use of sentiment dictionaries such as the Loughran-McDonald Dictionary, which have been specifically tailored for financial text [4]. Rule-based systems are generally straightforward to implement and can perform well on texts that closely match the predefined rules. However, they often struggle with the complexity and variability of natural language, particularly in financial news, where context and subtle sentiment play crucial roles.

As the limitations of rule-based methods became apparent, traditional machine learning techniques emerged as a more flexible and robust solution for sentiment analysis. These methods typically involve training classifiers such as Naive Bayes, Support Vector Machines (SVM), and logistic regression on labeled datasets. These classifiers can learn to identify patterns and relationships between words and sentiment labels, allowing for more accurate sentiment predictions. For example, a study demonstrated the utility of using SVMs to analyze the tone of financial news [5]. Despite their improved performance over rule-based methods, traditional machine learning models often require extensive feature engineering and can struggle with capturing the intricate dependencies present in financial news.

The advent of deep learning has significantly advanced the field of sentiment analysis. Deep learning models, particularly recurrent neural networks (RNNs) [6], convolutional neural networks (CNNs) [7], and hybrid models like CNN-LSTM [8], have shown remarkable capabilities in automatically learning hierarchical features from raw text data. In the financial domain, researchers have leveraged these models to capture the complex syntactic and semantic structures of news articles. Long Short-Term Memory (LSTM) networks, a variant of RNN, have been particularly effective in modeling

sequential dependencies, which are crucial for understanding the context of financial news. For instance, a study [9] used LSTM to analyze financial news sentiment, demonstrating superior performance compared to other benchmark models.

Recent advancements in pre-trained language models (PLMs) such as BERT [10], GPT [11], GLM [12], and their variants have significantly advanced sentiment analysis. In particular, LLMs, trained on vast amounts of data, have become super powerful and can be fine-tuned for specific tasks, including sentiment analysis. Their ability to understand context and generate human-like text has opened new avenues for enhancing sentiment analysis in the financial sector. Leveraging LLMs, researchers have utilized prompt engineering methods [13] to embed financial news and social media text into pre-designed prompt templates, enabling these models to generate sentiment scores. Recent studies have shown that prompt engineering techniques can significantly improve the accuracy of sentiment analysis across various domains [14]. In the financial news sentiment analysis field, some studies [15, 16] have utilized LLMs to categorize sentiment into negative, neutral, or positive labels, which are then averaged across multiple news items or posts to generate a robust daily sentiment indicator. These sentiment scores are often used as input features for downstream financial time series prediction models, improving their accuracy. These methods are still in their pre-mature stage, and among the very limited related work in the literature, most attempts have adopted only zero-shot and few-shot strategies. Following this trend, our proposed DK-CoT strategy builds on this foundation by introducing a novel prompt design that further enhances the model's ability to interpret and predict market sentiment from news articles.

The development of fine-tuning techniques is in general closely associated with the advancement of the deep learning model itself. The advent of the earliest generation of fine-tuning techniques was in line with the increasing size of the deep learning model. The cost of training using the same model architecture, but a different dataset was excessive and there might be a deterioration of the model capability after training. Since then, the growing size of the model has been far more than the development of hardware performance, resulting in the emergence of more efficient fine-tuning techniques. The reason why more recent cutting-edge prompt strategies [17] have become increasingly prominent as an alternative to traditional fine-tuning methods is twofold. Firstly, fine-tuning large models such as LLMs is becoming increasingly impossible for common hardware settings. Secondly, the large model itself is sufficiently capable to make prompt-finetuning effective, so it makes more sense to leverage such capability rather than starting from scratch. This in turn contributes to maintaining sustainability from the perspective of the shared future of mankind.

In summary, the comparison of rule-based AI, traditional machine learning (ML), deep learning, and pre-trained models (such as BERT, RoBERTa and LLMs) is shown in Table 1. The LLMs together with prompting strategies provide a possibility of more accessible and reasonably interpretable AI with low-level pre-processing.

3 DK-CoT: the proposed prompt tuning strategy

LLMs are trained on large amounts of unlabeled text. Despite their excellent performance in NLP tasks, their generality leads to a wide range of responses with a certain degree of uncertainty, which may not be well-suited for specific scenarios or domains. However, the emergence of prompt engineering, a new discipline focused on the development and optimization of prompts, has played a crucial role in unlocking the potential of LLMs [18]. This approach, which requires the construction of prompt phrases, has a lower barrier to entry and is more suitable for individuals without information technology (IT) backgrounds [19]. Researchers design appropriate prompt templates to enhance the ability of LLMs to handle complex task scenarios, ensuring the accuracy, relevance, and consistency of the model's output in specific domains [20]. Prevalent prompt engineering techniques include zero-shot prompts, few-shot prompts, and CoT prompts. The impact of prompt engineering extends to various disciplines; for example, it has facilitated the creation of robust feature extractors using LLMs, thereby improving their efficacy in tasks such as defect detection and classification [21].

Table 1 Comparison of different sentiment analysis techniques

Dimension	Rule-based AI	Traditional ML	Deep learning	Pre-trained models
Complexity	Low	Medium	High	High
Pre-processing	High	High	Low	Low
Accessibility	Medium	Medium	Low	High
Interpretability	Good	Good	Bad	Varied

In this study, we have proposed a prompt engineering framework with a novel prompting strategy called DK-CoT, which is a hybrid approach combining the insertion of financial domain knowledge, as well as CoT reasoning, aiming to enhance the performance in the task of financial news sentiment classification. To evaluate its performance, we compare it with various other prompt strategies, including zero-shot, few-shot, CoT, and generative knowledge, as well as some state-of-the-art sentiment analysis models, including BERT and RoBERTa. Figure 2 illustrates the framework, where the source and the retrieval of the finance domain knowledge can be flexible, such as finance domain experts, knowledge bases, etc. The knowledge retrieval can be directly from the established knowledge base by domain experts or leveraging more advanced methods like Retrieval-Augmented Generation (RAG) for the full automation of knowledge retrieval and LLM output. In the experiment of this study, domain knowledge for each target company to be analyzed is provided by finance domain experts, saved in a knowledge base, and retrieved based on the name of the target company, which can be considered a light version of RAG. The LLM we use in this study is GLM, which will be illustrated in Sect. 3.2.

The rest of this section will illustrate the relevant techniques involved in this study.

3.1 State-of-the-art Models: BERT and RoBERTa

BERT is a pre-trained language model that has achieved significant success in the NLP field. It primarily employs the Encoder from the Transformer model, which consists of multiple layers of self-attention and feedforward neural network layers, suitable for sequential data. The BERT model mainly learns feature representations of the input data and applies these learned features to downstream tasks. Traditional deep learning models typically use unidirectional structures like left-to-right or right-to-left. This structural design limits the performance of certain downstream tasks. However, by adopting a bidirectional structural design, BERT can read in both directions, resulting in better performance.

RoBERTa, an enhanced version of BERT, brings improvements and optimizations in training, resulting in better performance on various NLP tasks. RoBERTa introduces four key enhancements on top of BERT, including improved

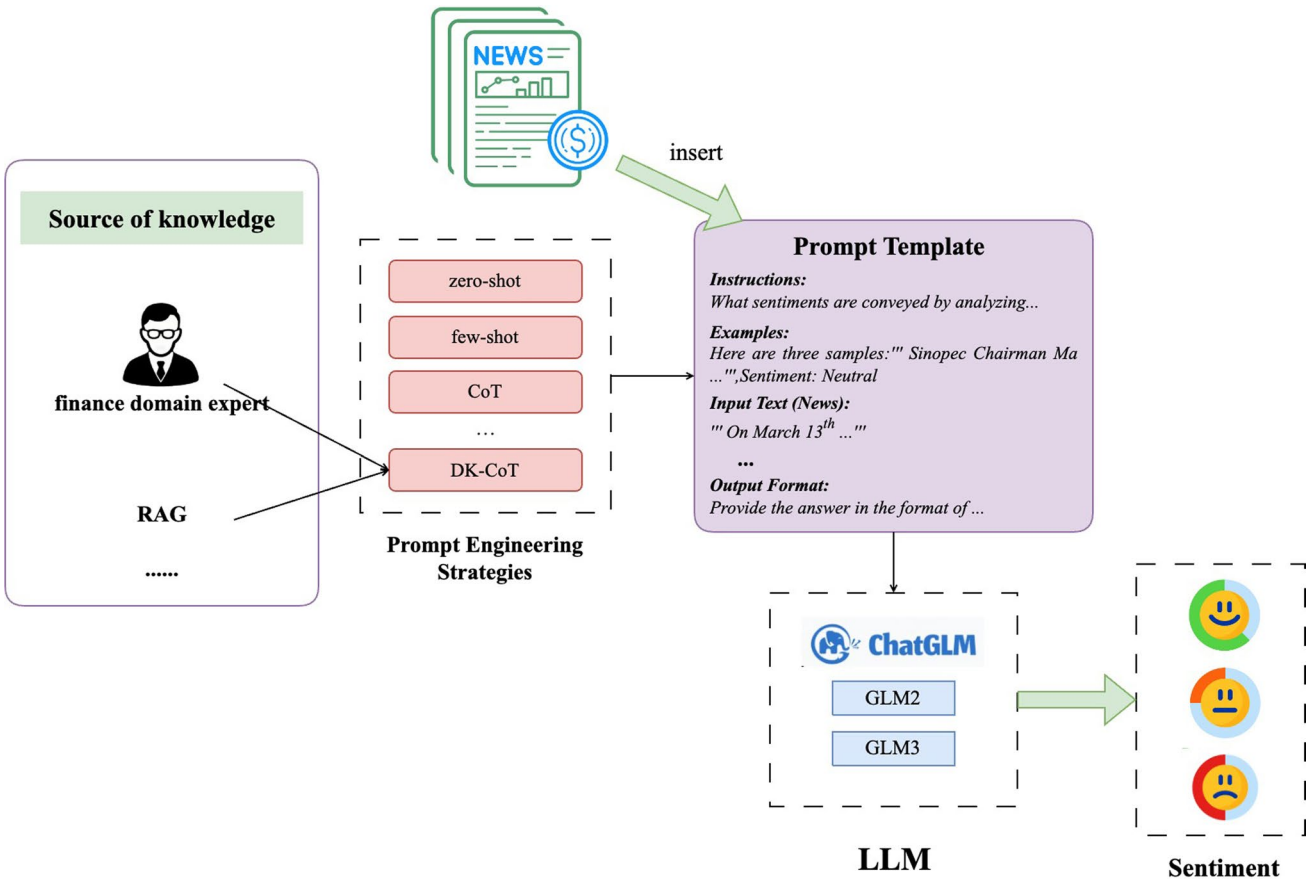


Fig. 2 The framework for LLM-enhanced financial news sentiment analysis

masking strategy, removal of the next sentence prediction task, the use of larger training data and batch sizes, and changes in text encoding.

BERT and RoBERTa have set the state-of-the-art benchmark for NLP tasks, including text classification, named entity recognition, and sentiment analysis. Specifically, in the domain of financial news sentiment analysis, these models have been widely used due to their ability to capture contextual dependencies and nuanced meanings, which are crucial for accurately interpreting financial text. There are also some variations of BERT or RoBERTa that are fine-tuned using financial data, e.g. FinBERT [22] and FLANG-RoBERTa [23], which have shown superior results in financial news sentiment analysis.

For instance, one study applied a deep learning approach using the BERT model on a financial market dataset, achieving high performance with an accuracy of 95.29% and an F1-score of 95.32%, highlighting BERT's effectiveness in conducting sentiment analysis in financial markets [24]. Another study compared sentiment analysis models and found that RoBERTa and FinBERT achieve the highest average accuracy and F1 score, confirming the effectiveness of BERT, RoBERTa and their variations [25]. Kirtac and Germano demonstrated that BERT and RoBERTa significantly enhance financial sentiment analysis and trading strategy development, highlighting their effectiveness in processing extensive financial news datasets [26]. Given their strong performance in general sentiment classification tasks, we adopt BERT, RoBERTa, and their variations including FinBERT and FLANG-RoBERTa as baselines to evaluate the effectiveness of our proposed method, DK-CoT.

3.2 GLM

Pre-trained language models are primarily divided into three categories: autoregressive models, autoencoder models, and encoder-decoder models. Each of these models has its advantages and disadvantages, and no single model currently excels in all NLP tasks. Some researchers have attempted to combine different frameworks, but due to the inherent differences between autoencoding and autoregressive objectives, simple combinations cannot fully exploit their respective strengths. With the advent of ChatGPT, breakthroughs and challenges have emerged for NLP tasks. ChatGPT is an LLM based on the GPT architecture, demonstrating outstanding performance in various NLP tasks, including text summarization, text completion, language translation, and question answering [27]. However, despite the remarkable achievements of ChatGPT in the field of NLP, its developer OpenAI has not publicly released the model's source code and its weights. This limits other institutions and enterprises in their ability to study and apply the model. In response, Zhipu AI,[1] powered by the technological advancements of Tsinghua University, has released a GPT equivalent called GLM (General Language Model) and has made some base models open source [28]. Note that GLM is the name of the model, but the chatbot based on this model is called ChatGLM, akin to the relationship between GPT and ChatGPT. GLM adopts the Transformer-based architecture and self-attention mechanism. It employs a mode-exploiting training approach that converts natural language understanding tasks into word cloze tasks with task descriptions. This involves randomly removing consecutive text segments from the input text and training the model to restore these segments in a certain order. This method combines the advantages of both autoregressive and autoencoding pre-training approaches. Additionally, GLM incorporates two-dimensional positional encoding and allows for arbitrary order prediction of blank regions, improving the blank-filling pre-training process. Users can deploy and customize GLM, which lowers the barriers to academic research and corporate development of LLMs and promotes further development and application of NLP technology.

In this study, we employ GLM2-6B and GLM3-6B models, referred to as GLM2 and GLM3 in the rest of the paper. The reasons why we chose GLM as the LLM for this study are three-fold. Firstly, GLM has been trained using the most up-to-date English and Chinese texts and has been proven to perform well on many metrics [3], making it one of the most suitable models considering our study focuses on a Chinese context. Secondly, the weights of the GLM-6B series models are fully open for academic research and allow free commercial use. Lastly, GLM can be deployed locally, supports model quantization, and can be deployed on CPUs. Section 3.3 will illustrate the five prompting strategies we have applied to GLM, including the proposed DK-CoT.

[1] https://www.zhipuai.cn/en/

3.3 Prompting Strategies

This section illustrates the prompting strategies involved in this study and how the prompts are formed. The prompts are formed incrementally by adding elements from zero-shot, few-shot, and all the way to our proposed DK-CoT.

3.3.1 Zero-shot prompting

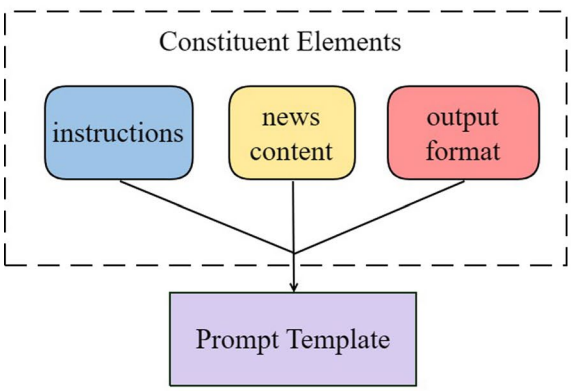
Zero-shot prompting [29] is a technique that designs a simple prompt template with specific instructions, which leverages the inherent capabilities and the pre-trained knowledge of the LLM to perform tasks it was not explicitly trained on without the need to fine-tune the LLM [30]. Thus, this method can be considered as a form of transfer learning. In the experiment, we provided GLM with a series of texts, instructing the model on what to do using natural language descriptions. The zero-shot prompt template used in this study is shown in Fig. 3. Specifically, the zero-shot prompt template consists of instructions on the task to be conducted, the news content to be analyzed, and guidance on the output format.

3.3.2 Few-shot prompting

LLMs with zero-shot prompting may underperform on certain complex tasks, showing limitations in their capabilities. To address this issue, few-shot prompting can be employed. In this approach, the prompt template provides the LLM with some examples similar to the task to be performed, requiring the model to answer questions the way shown in these examples [31]. Compared with zero-shot prompting, few-shot prompting reduces the dependency on the dataset for the specific task. The few-shot prompt template used in this study is shown in Fig. 4, where one example of financial news for each type of sentiment label (positive, neutral, and negative) is provided for GLM to learn.

3.3.3 Chain-of-thought (CoT) prompting

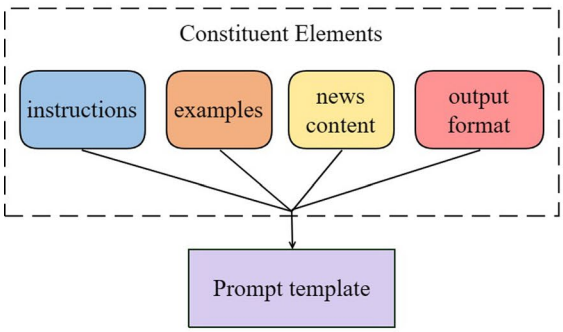
Even with zero-shot and few-shot prompting techniques, LLMs may struggle to produce reliable responses when handling complex reasoning tasks. Increasing the model size and the number of parameters alone is insufficient to address this challenge [32], hence the advent of the chain-of-thought (CoT) prompting technique [33]. This technique can be combined with few-shot prompting, where intermediate steps of computation and reasoning are mimicked using examples before tackling complex tasks, thereby improving the results. The CoT prompt template used in this study is shown in Fig. 5.



Prompt(Chinese): 分析三引号中的新闻内容传达了什么情绪？是正面、负面还是中性的？如果无法判断新闻内容表达的情绪，默认答案为中性。按照【情绪：正面】的格式给出答案，不要给出解释。''' 新闻内容 '''，情绪：
Response(Chinese): 正面/中性/负面

Prompt(English): What sentiments are conveyed by analyzing the news content within three quotation marks? Is it positive, negative, or neutral? If it is impossible to determine the sentiment expressed in the news content, the default answer is neutral. Provide the answer in the format of [sentiment: Positive], without providing the explanation. '''News Full Text''' sentiment:
Response(English): Positive/Neutral/Negative

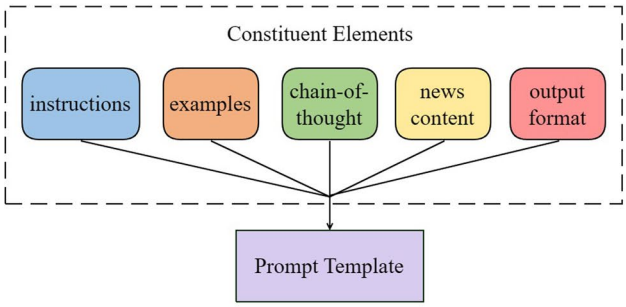
Fig. 3 Zero-shot prompt template



Prompt(Chinese): 以下是三个样本：'''3月25日，中国石化（600028）举行2023年度业绩说明会...''', 情绪：中性；'''证券之星消息，中国石化2023年年报显示...''', 回答：负面；'''3月13日，中国石化与宁德时代新能源科技股份有限公司在北京签署战略合作框架协议...''', 情绪：正面。参考上述样本案例，分析三引号中的新闻内容传达了什么情绪？是正面、负面还是中性的？如果无法判断新闻内容表达的情绪，默认答案为中性。按照【情绪：正面】的格式给出答案，不要给出解释。''' 新闻内容 ''', 情绪：
Response(Chinese): 正面/中性/负面

Prompt(English): Here are three samples:'''On March 25, Sinopec (600028) held its 2023 annual performance briefing...''', Sentiment: Neutral;'''According to Securities Times, the 2023 annual report of Sinopec shows...''', Sentiment: Negative;'''On March 13, Sinopec and Contemporary Amperex Technology Co., Limited signed a strategic cooperation framework agreement in Beijing...''', Sentiment: Positive. Referring to the above sample cases, please analyze what sentiment is conveyed by the news content within three quotation marks? Is it positive, negative, or neutral? If it is impossible to determine the sentiment expressed in the news content, the default answer is neutral. Provide the answer in the format of [sentiment: Positive], without providing the explanation. '''News Full Text''' sentiment:
Response(English): Positive/Neutral/Negative

Fig. 4 Few-shot prompt template



Prompt(Chinese): 以下是三个样本：'''3月25日，中国石化（600028）举行2023年度业绩说明会...''', 原因分析：虽然新闻中提到了中国石化...。情绪：中性；'''证券之星消息，中国石化2023年年报显示...''', 原因分析：主营收入和净利润都出现了同比下降...。情绪：负面；'''3月13日，中国石化与宁德时代新能源科技股份有限公司在北京签署战略合作框架协议...''', 原因分析：这种增持表明南向资金...。情绪：正面。参考上述样本案例，分析三引号中的新闻内容传达了什么情绪？是正面、负面还是中性的？如果无法判断新闻内容表达的情绪，默认答案为中性。按照【情绪：正面】的格式给出答案，不要给出解释。''' 新闻内容 ''', 情绪：
Response(Chinese): 正面/中性/负面

Prompt(English): Here are three samples:'''On March 25, Sinopec (600028) held its 2023 annual performance briefing...''' Reason analysis: although the news mentioned Sinopec... sentiment: Neutral;'''According to Securities Times, the 2023 annual report of Sinopec shows...''' Reason analysis: both main revenue and net profit have decreased year-on-year...; Sentiment: Negative;'''On March 13, Sinopec and Contemporary Amperex Technology Co., Limited signed a strategic cooperation framework agreement in Beijing...''' Reason analysis: this increase in holdings indicates that southbound funds... ; Sentiment: Positive. Referring to the above sample cases, please analyze the sentiment is conveyed by the news content within three quotation marks? Is it positive, negative, or neutral? If it is impossible to determine the sentiments expressed in the news content, the default answer is neutral. Provide the answer in the format of [sentiment: Positive], without providing the explanation. '''News Full Text''' sentiment:
Response(English): Positive/Neutral/Negative

Fig. 5 Chain-of-thought prompt template

3.3.4 Generative knowledge prompting

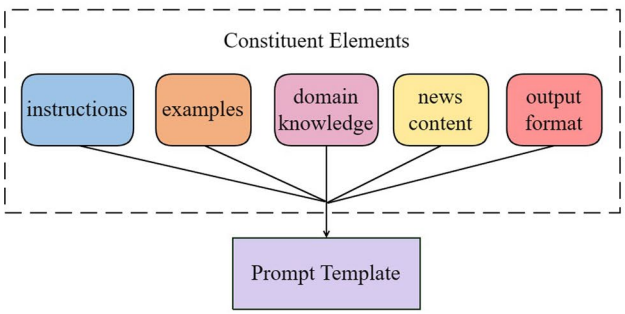
In certain specialized domains, LLM may face additional challenges beyond CoT prompting. For example, due to a lack of exposure to specific domain datasets or a lack of relevant knowledge in that domain during the training process, even if CoT prompts are introduced in few-shot prompting techniques, the model may perform poorly, which limits its ability to effectively transfer learning. To address this issue, generative knowledge prompting techniques can be

employed [34]. This method retains the model's flexibility while guiding it to generate domain-specific knowledge and then connecting this knowledge to the problem as input.

In this section, we combine few-shot prompting techniques with generative knowledge prompting techniques to provide GLM with a background in the financial domain and design an appropriate prompt template to enhance its performance in financial sentiment classification tasks. We use the demonstration examples from Sect. 3.3.2 for few-shot prompting. For generative knowledge prompting, we only utilize the second step of the technique, providing GLM with financial domain knowledge and connecting this knowledge to the problem as input to GLM. The combined few-shot and generative knowledge prompt template for the financial news sentiment classification task is shown in Fig. 6.

3.3.5 DK-CoT prompting strategy

While zero-shot prompting, few-shot prompting, generative knowledge prompting, and CoT prompting each improve LLM performance to varying degrees, LLMs are typically trained on historical datasets, limiting their ability to learn and adapt to new data, which can lead to suboptimal performance. Additionally, these prompting techniques focus on enhancing LLM performance based on a single factor, overlooking the influence of other factors. For instance, CoT prompting improves model performance by leveraging computation and reasoning, but it may overlook the importance of incorporating domain knowledge. Conversely, generative knowledge prompting focuses on integrating domain knowledge but does not involve computation and reasoning processes. As a result, a particular prompting technique may offer the best performance enhancement in a specific domain but may be outperformed by other techniques in different domains, failing to ensure consistent model performance. To address this, DK-CoT retains the advantages of these prompting techniques while mitigating their shortcomings, ensuring excellent model performance across different domains. DK-CoT provides domain knowledge within the prompt template and dynamically adjusts based on current domain knowledge changes, overcoming the limitations of LLMs being restricted to historical datasets and allowing them to learn the latest knowledge promptly. This method also integrates the intermediate computation and reasoning processes of CoT prompting, incorporating the most recent domain knowledge to ensure LLMs can provide more accurate and up-to-date responses in practical applications, significantly enhancing performance and stability. The prompt template designed using DK-CoT, incorporating



Prompt(Chinese): 已知金融新闻对股票市场的影响是复杂而多样的。新闻中的积极或消极或中性信息会...;以下是三个样本：'''3月25日，中国石化（600028）举行2023年度业绩说明会...'''，情绪：中性；'''证券之星消息，中国石化2023年年报显示...'''，回答：负面；'''3月13日，中国石化与宁德时代新能源科技股份有限公司在北京签署战略合作框架协议...'''，情绪：正面。参考上述样本案例，结合提供的领域知识分析三引号中的新闻内容传达了什么情绪？是正面、负面还是中性的？如果无法判断新闻内容表达的情绪，默认答案为中性。按照【情绪：正面】的格式给出答案，不要给出解释。
''' 新闻内容 ''' 情绪：
Response(Chinese): 正面/中性/负面

Prompt(English): The impact of financial news on the stock market is known to be complex and diverse. Positive, negative, or neutral information in news can...;Here are three samples:'''On March 25, Sinopec (600028) held its 2023 annual performance briefing...''', Sentiment: Neutral;'''According to Securities Times, the 2023 annual report of Sinopec shows...''', Sentiment: Negative;'''On March 13, Sinopec and Contemporary Amperex Technology Co., Limited signed a strategic cooperation framework agreement in Beijing...''', Sentiment: Positive. Referring to the above sample cases, please analyze what emotion is conveyed by the news content in triple quotes based on the provided domain knowledge. Is it positive, negative, or neutral? If it is impossible to determine the sentiments expressed in the news content, the default answer is neutral. Provide the answer in the format of [sentiment: Positive], without providing the explanation. '''News Full Text''', sentiment:
Response(English): Positive/Neutral/Negative

Fig. 6 Generative knowledge prompting

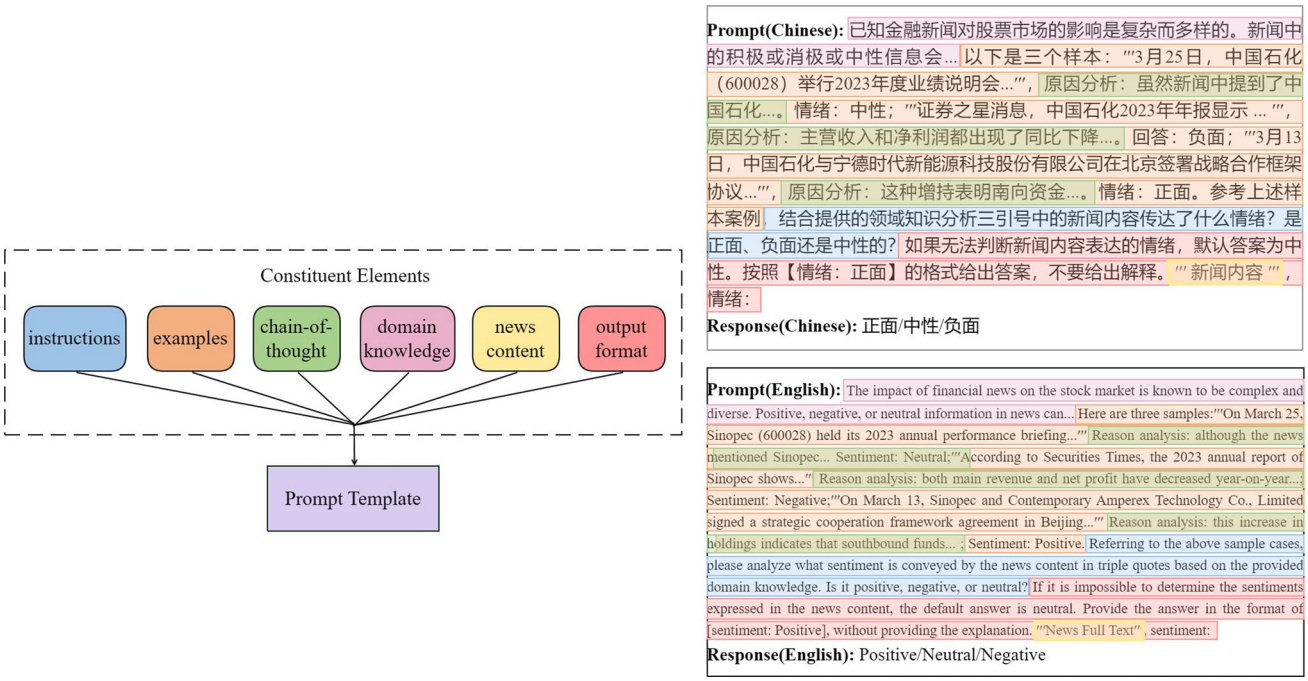


Fig. 7 DK-CoT prompting strategy

domain knowledge, examples, and reasoning processes, is shown in Fig. 7. Two full DK-CoT prompt examples for two different companies in both Chinese and English have been provided in the Appendix.

4 Experimental setup

4.1 Data and prompt formation

The data for this study was sourced from the Baidu Stock Market[2] platform, which has pre-labeled each financial news article with sentiment tags. The datasets used in the experiment consist of 288 news articles for Sinopec (from February 22, 2024, to March 28, 2024, and from June 19, 2024, to July 27, 2024), 130 for PetroChina (from April 4, 2024, to April 23, 2024), 549 for Xiaomi Group (referred to as "MI" for the rest of the paper, from May 24, 2024, to July 29, 2024), 120 for YunXingYu (referred to as "BYTTL" for the rest of the paper, from January 12, 2024, to July 26, 2024), and 957 of "data integrated" (referred to as "DI" for the rest of the paper, a combination of the datasets from Sinopec, Xiaomi Group, and YunXingYu). The datasets primarily include two features, i.e. the full content of the news articles and the sentiment labels (positive, negative, and neutral). The sentiment labels were mapped as follows in the code: positive as 1, negative as -1, and neutral as 0.

The distribution of data types for each dataset is shown in Fig. 8. It should be noted that in the context of financial news, neutral sentiments are naturally more prevalent than positive or negative sentiments. Financial news articles often report factual information, market updates, and analysis without imparting a strong positive or negative sentiment. This real-world distribution is essential for evaluating if a model performs well under actual operating conditions. If we artificially balance the dataset, the model might perform well on a balanced test set but fail to generalize to real-world data where neutral sentiments dominate.

Among these datasets, the Sinopec dataset (288 news articles) is used to explore effective prompt templates, whereas the PetroChina dataset (130 news articles) is used for model fine-tuning (where applicable). Sinopec, Xiaomi Group (549 articles), and YunXingYu (120 articles), totaling 957 articles were used to test the effectiveness and stability of the DK-CoT prompting strategy.

[2] https://gushitong.baidu.com/

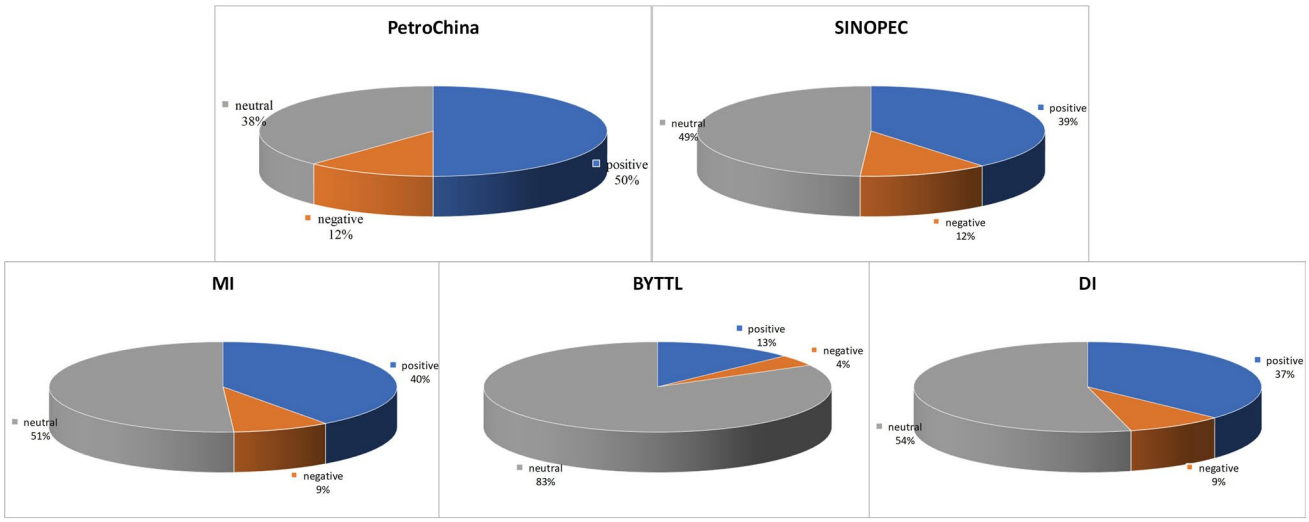


Fig. 8 Datasets and distribution

In the experiment, for prompt templates involving examples, three news articles with different sentiment labels were uniformly inserted into the prompt templates, ensuring that the model had reference samples for accurate classification of the remaining news articles. For templates involving domain knowledge, a finance domain expert has been consulted, and the domain knowledge for each company added to the prompt templates has been sourced from the knowledge base established via the domain expert, including but not limited to background information about the relevant companies, competitors, recent announcements, and the impact of financing activities on the companies. This supplements the model's lack of knowledge, especially after the training of the model. This comprehensive domain knowledge is crucial for achieving accurate sentiment classification.

4.2 Hardware and Software Settings

The experiment was conducted on a 64-bit operating system with an × 64-based processor (13th generation Intel (R) Core (TM) i5-13600KF 3.50 GHz), 32.0 GB of RAM, and an NVIDIA GeForce RTX 4070 Ti (12282 MB) graphics card. All models are implemented in Python 3.8 and PyTorch.

The experiment involves GLM, BERT, fine-tuned BERT (FinBERT), FLANG-RoBERTa, and fine-tuned FLANG-RoBERTa (Fin-FLANG-RoBERTa) models. For GLM, various prompt techniques such as zero-shot, few-shot, CoT, generative knowledge, and DK-CoT are employed to design prompt templates, which are then tested on the test set. Similarly, BERT (bert-base-chinese), FinBERT (yiyanghkust/finbert-tone-chinese), and FLANG-RoBERTa (chinese-roberta-wwm-ext-FineTuned) are also tested on the test set. The vanilla FLANG-RoBERTa performed poorly so we also fine-tuned FLANG ROBERTa using our training set to obtain Fin-FLANG-RoBERTa, which was then tested on the testing set. Due to the maximum length limitation of FLANG-RoBERTa models, sentences exceeding 512 tokens are processed by truncating the first 128 tokens and the last 382 tokens.

4.3 Evaluation metrics

This experiment utilizes commonly used evaluation metrics for sentiment classification, such as recall, precision, accuracy, and F1 score, which are calculated using the following formulas:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\_Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

We have also introduced the weighted F1 score as an additional evaluation metric. The rationale is that different sentiments have varying levels of attention and impact on stock market dynamics. Negative news is generally considered the most impactful, spreading quickly and widely, and attracting more attention. Behavioral finance suggests that humans are more sensitive to losses than gains, meaning the pain from a loss is greater than the pleasure from an equivalent gain. This loss aversion leads investors to react more strongly to negative news, often causing panic selling and market volatility [35]. In contrast, positive news, while influential, elicits more rational and moderate reactions, usually having a less intense impact. Neutral news, with lower information density and no clear trading signals, attracts less attention and does not significantly influence market volatility. Based on consultations with finance domain experts, it is recommended to assign different weights to negative sentiment, positive sentiment, and neutral sentiment according to their varying degrees of impact on market volatility. These weights are 0.5, 0.3, and 0.2, respectively.

First, calculate the precision and recall for all three categories, for Category i:

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

Let weights = [0.5,0.3,0.2], and calculate the weighted precision and weighted recall accordingly:

$$Weighted\_Precision = \sum_{i=1}^{3} weight_i * Precision_i$$

$$Weighted\_Recall = \sum_{i=1}^{3} weight_i * Recall_i$$

Finally, weighted F1 can be calculated using the following equation:

$$Weighted\_F1 = \frac{2 * Weighted\_Precision * Weighted\_Recall}{Weighted\_Precision + Weighted\_Recall}$$

The weighted F1 score can better reflect the actual impact of different sentiment categories on the financial market, enhancing the effectiveness of sentiment classification models in practical applications. This method integrates traditional sentiment analysis techniques with the behavioral characteristics of the financial market, making the sentiment classification results more relevant and actionable, particularly when followed by market dynamics analysis like event studies and stock price predictions [36], thereby providing more reliable support for investment decisions and risk management.

5 Results and discussion

5.1 Experimental results

The models involved in this study include BERT, fine-tuned BERT (FinBERT), FLANG-RoBERTa, fine-tuned FLANG-RoBERTa (Fin-FLANG-RoBERTa), GLM2, and GLM3. The prompting techniques discussed in Sect. 3.3 have been

applied to both GLM2 and GLM3. Considering the uncertainty in the generated content by GLM, the sentiment category for each news item was determined based on the voting results from twenty rounds of executions for each prompt template. The rest of this section will demonstrate the experimental results for each model in detail.

5.1.1 BERT, FinBERT, FLANG-RoBERTa, and Fin-FLANG-RoBERTa

Figure 9 shows the experimental results of BERT and FinBERT on the Sinopec dataset. It can be observed that BERT demonstrates good ability to distinguish neutral and positive sentiments, with recognition rates of 74% and 69%, respectively, but the recognition accuracy in predicting "negative/bearish" sentiments is poor, with a recognition rate of only 25%. Similarly, FinBERT performs well in distinguishing between neutral and positive sentiments, with recognition rates of 74% and 76%, respectively. However, its accuracy in predicting "negative/bearish" sentiments is very low, with a recognition rate of only 14%. It is not difficult to see from Fig. 9 (c) that all indicators of FinBERT have improved and are higher than BERT.

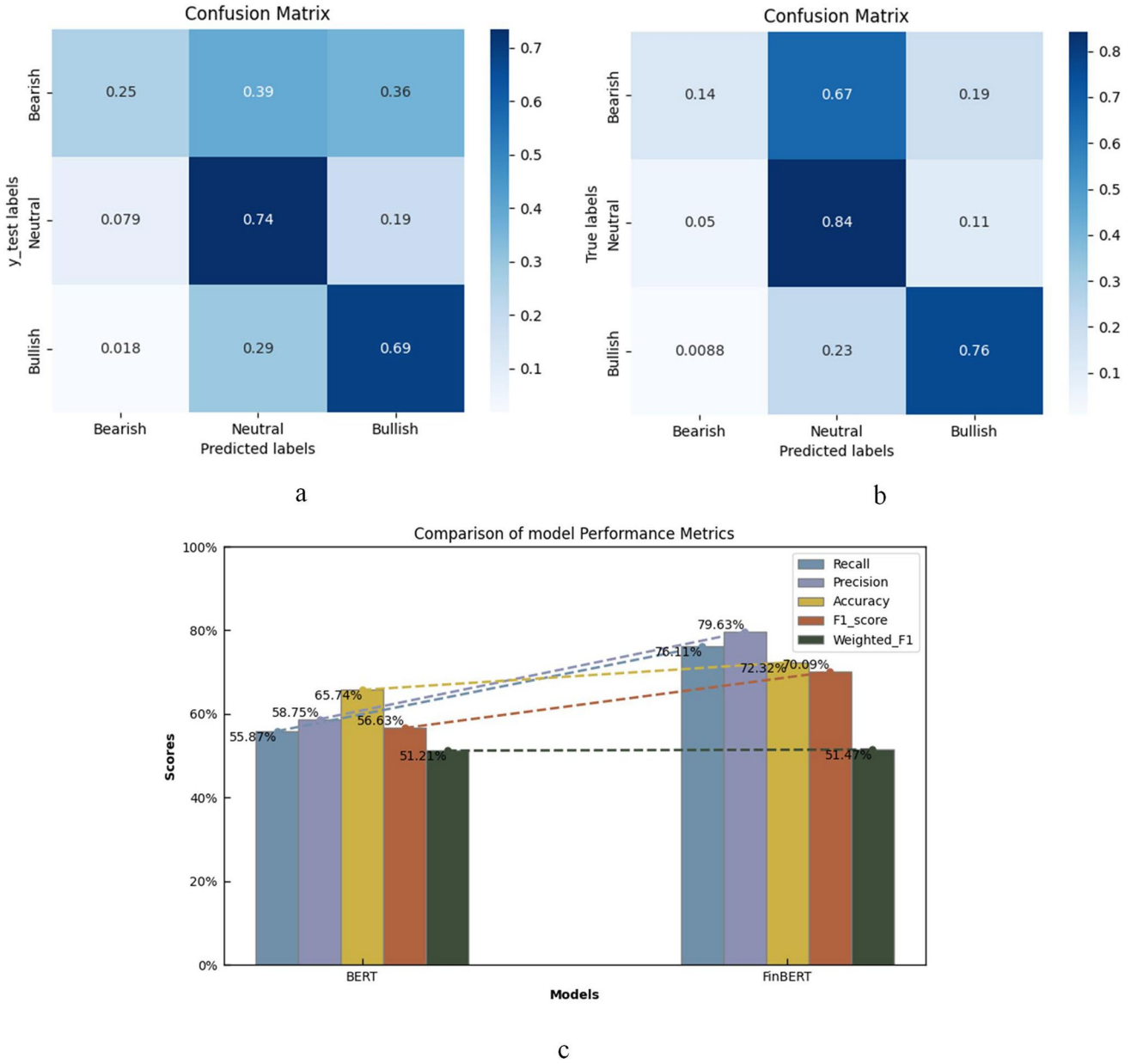


Fig. 9 Confusion matrix for **a** BERT, **b** FinBERT, and **c** their performance comparison

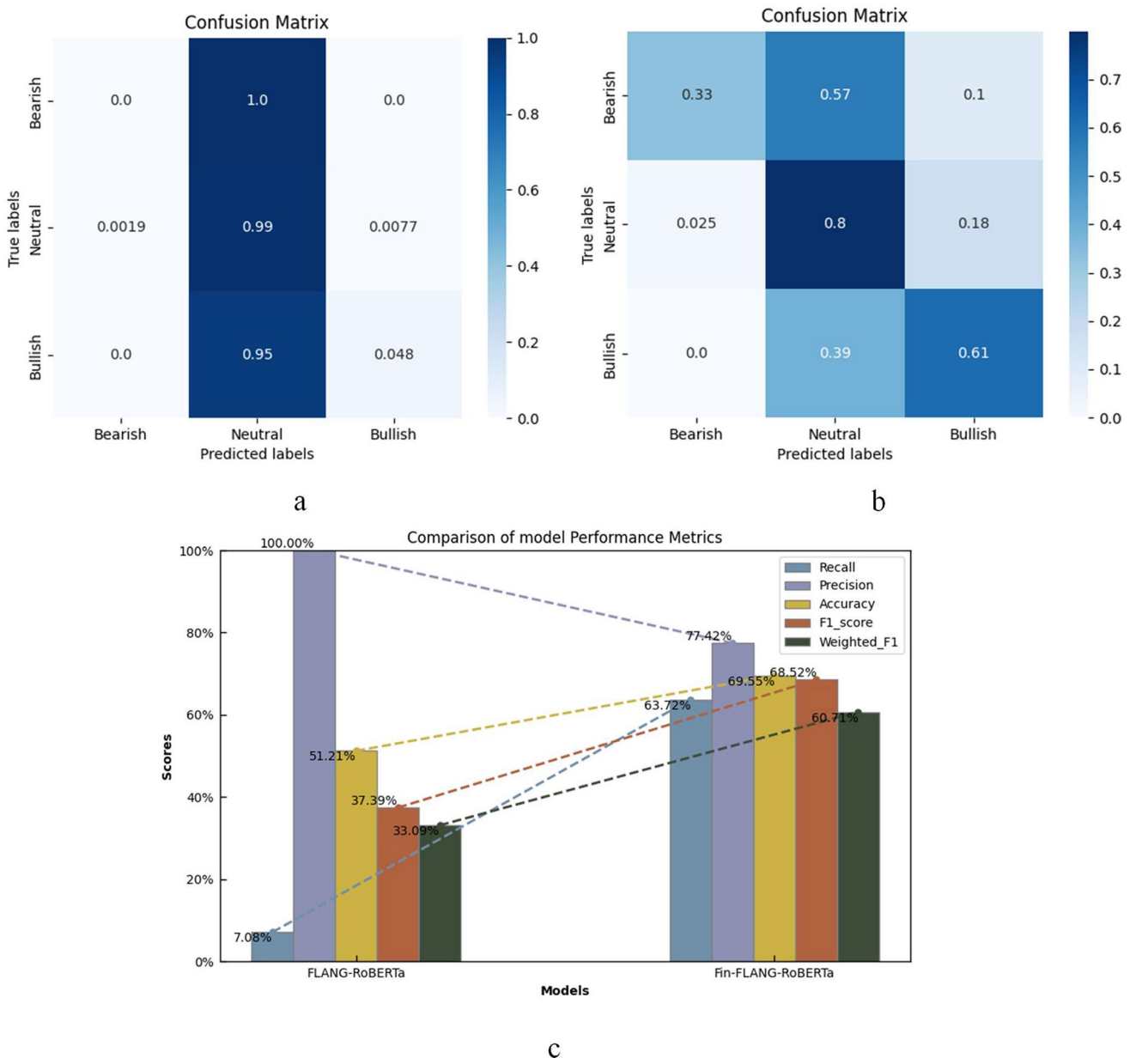


Fig. 10 Confusion matrix for **a** FLANG- RoBERTa, **b** Fin-FLANG-RoBERTa, and **c** their performance comparison

Figure 10 shows the experimental results of FLANG-RoBERTa and the further fine-tuned Fin-FLANG-RoBERTa. Without fine-tuning, FLANG-RoBERTa performs poorly, classifying the sentiment of almost all news as neutral, which is meaningless in practice. After fine-tuning using the Chinese petroleum dataset, the predictive performance of the Fin-FLANG-RoBERTa model improves significantly, especially for neutral and positive/bullish sentiment, while for negative/bearish sentiment, the accuracy remains low at only 33%. Considering negative news normally has a more profound impact, the inability to accurately identify negative news is still not desirable. It should be noted that there is a significant performance gap between the vanilla FLANG-RoBERTa and the fine-tuned FLANG-RoBERTa in our study, which can be attributed to the domain-specific nature of financial news sentiment analysis and the complexity of financial language. Vanilla FLANG-RoBERTa is pre-trained on general financial keywords and general corpora and lacks the understanding required for accurately interpreting specific financial news for a company. Financial news often contains specialized terminology and context-specific sentiments that are not prevalent in

general financial language usage. Fine-tuning FLANG-RoBERTa on a domain-specific dataset allows the model to learn these unique patterns, thereby substantially improving its performance. Thus, the significant performance gap highlights the importance of domain-specific fine-tuning in achieving high accuracy and reliability in specialized tasks like financial news sentiment analysis.

5.1.2 Prompting strategies on GLM2 and GLM3

The experimental results of zero-shot prompting for both GLM2 and GLM3 are shown in Fig. 11. It can be observed that GLM2 performs poorly on negative sentiments while performing well on positive and neutral ones. Conversely, GLM3 demonstrates better identification of negative and positive sentiments. GLM3 wins over GLM2 on all evaluation metrics when using zero-shot prompting, indicating superior performance.

The experimental results for the few-shot prompting strategy for GLM2 and GLM3 are shown in Fig. 12. Both GLM2 and GLM3 with few-shot prompting demonstrate better sentiment classification results of positive, negative, and neutral sentiments than those with zero-shot prompting. GLM3 outperforms GLM2 on all evaluation metrics except for a marginal lower precision for GLM3.

The experimental results of CoT for both GLM2 and GLM3 are described in Fig. 13. GLM2 performs well on neutral and positive sentiments, but poorly on negative sentiments. GLM3 exhibits good results for negative and positive sentiments, but not as good for neutral sentiments. Noticeably, GLM2 has caught up in terms of performance metrics under CoT, with almost as good results as seen in GLM3, indicating that CoT may enhance lower-level LLMs.

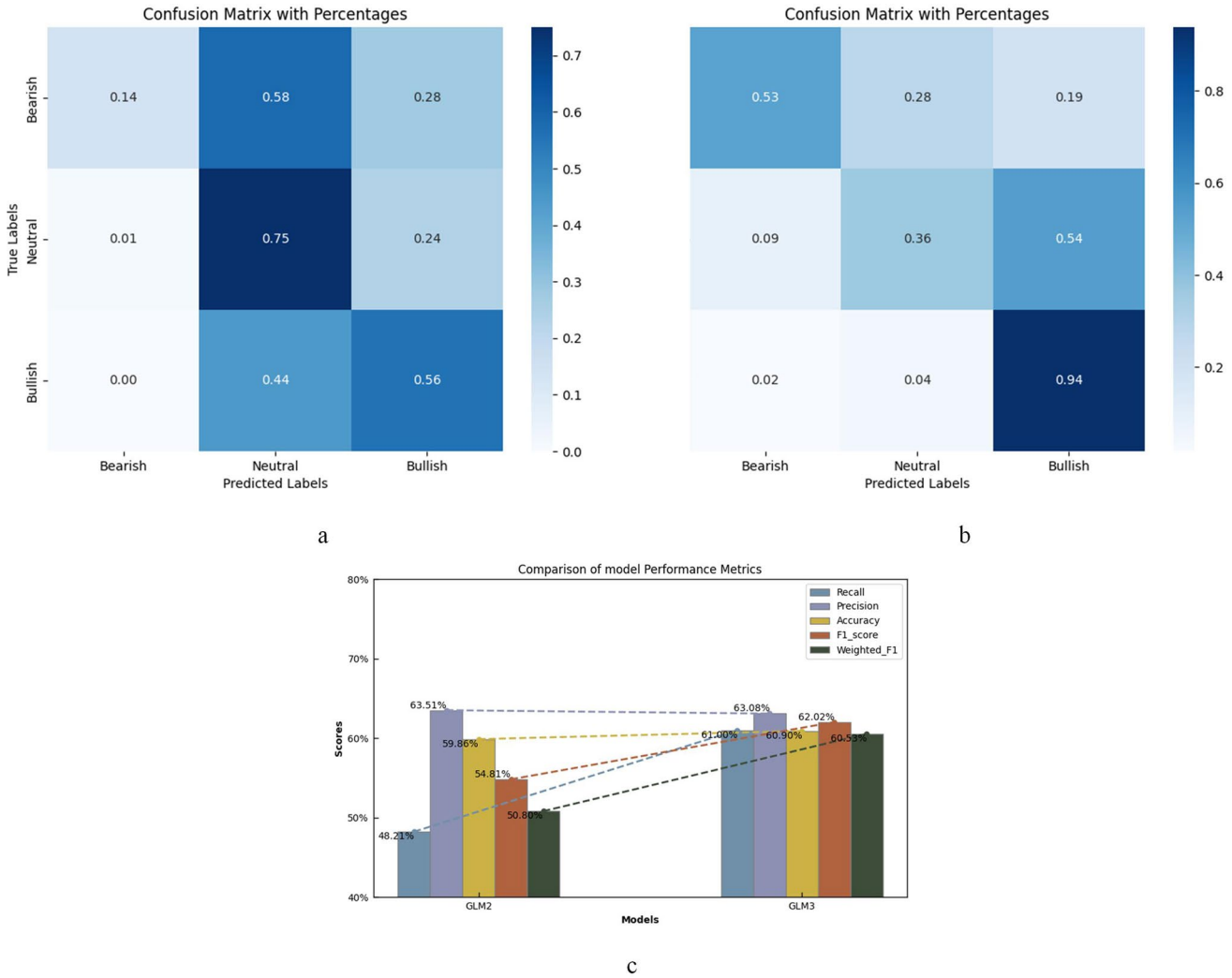


Fig. 11 Confusion matrix for zero-shot prompting on **a** GLM2 and **b** GLM3, and **c** their performance comparison

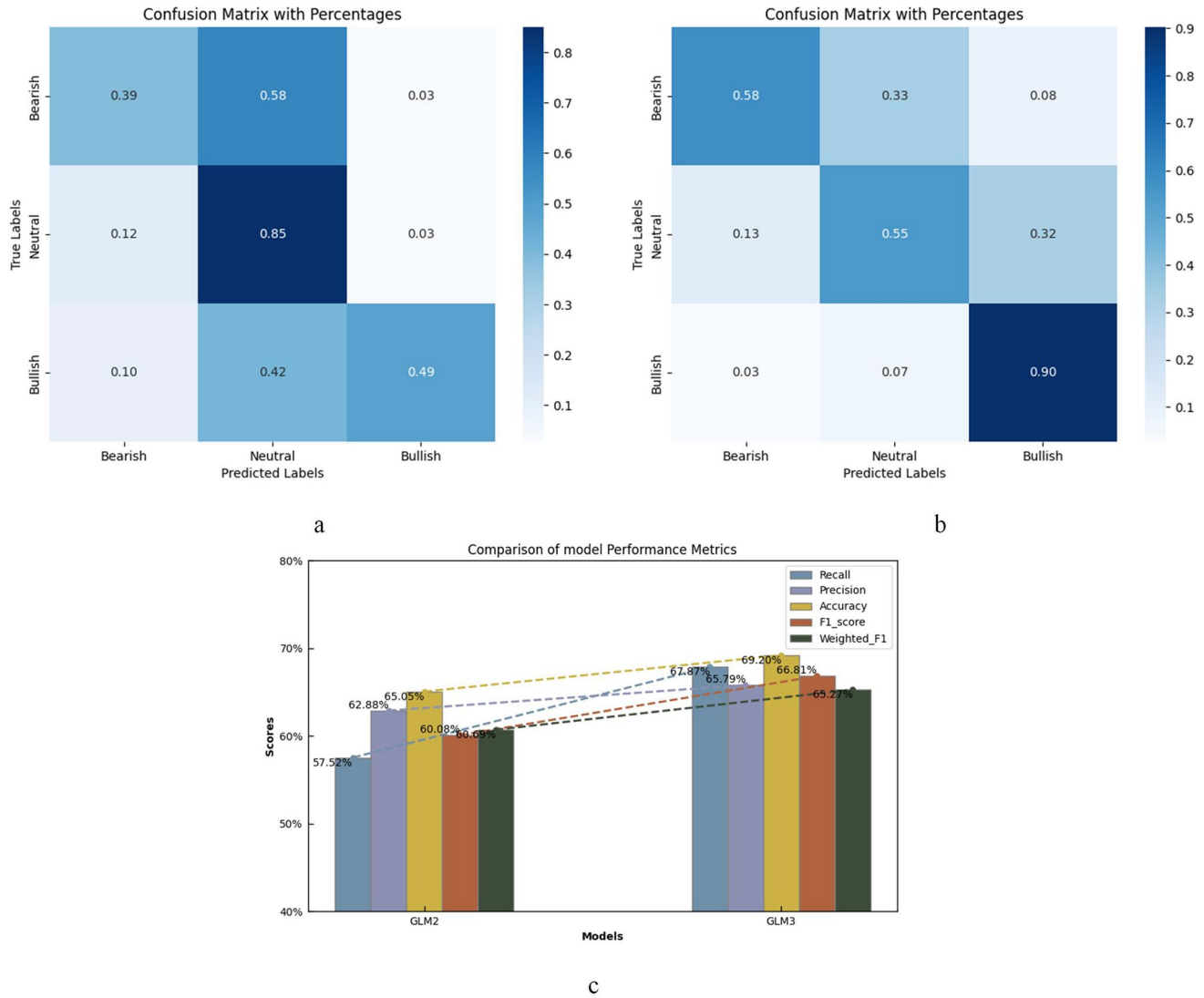


Fig. 12 Confusion matrix for few-shot prompting on **a** GLM2, **b** GLM3, and **c** their performance comparison

The experimental results for generative knowledge prompting for both GLM2 and GLM3 are shown in Fig. 14. GLM2 identifies most of the news items as "neutral", whereas GLM3 performs well on negative and positive sentiments, but not as well for neutral sentiments.

For our proposed strategy, DK-CoT prompting, Fig. 15 shows the experimental results. GLM2 follows a similar pattern as seen in other strategies, performing poorly on negative and positive sentiments, as it gives a "neutral" output for almost all news articles. However, GLM3 demonstrates better results for all three classes of sentiments with the highest weighted F1 score, indicating the strategy's effectiveness in balancing across all the categories.

5.2 Discussion, exploration, and validation

5.2.1 Findings

There are some important considerations in the DK-CoT strategy implementation. First, to mitigate ambiguity, we incorporated domain-specific knowledge directly into the prompt design, which helps the model disambiguate financial terminology and sentiment expressions based on context. For computational efficiency, our method leverages prompting without extensive fine-tuning, thus reducing resource consumption compared to traditional model adaptation

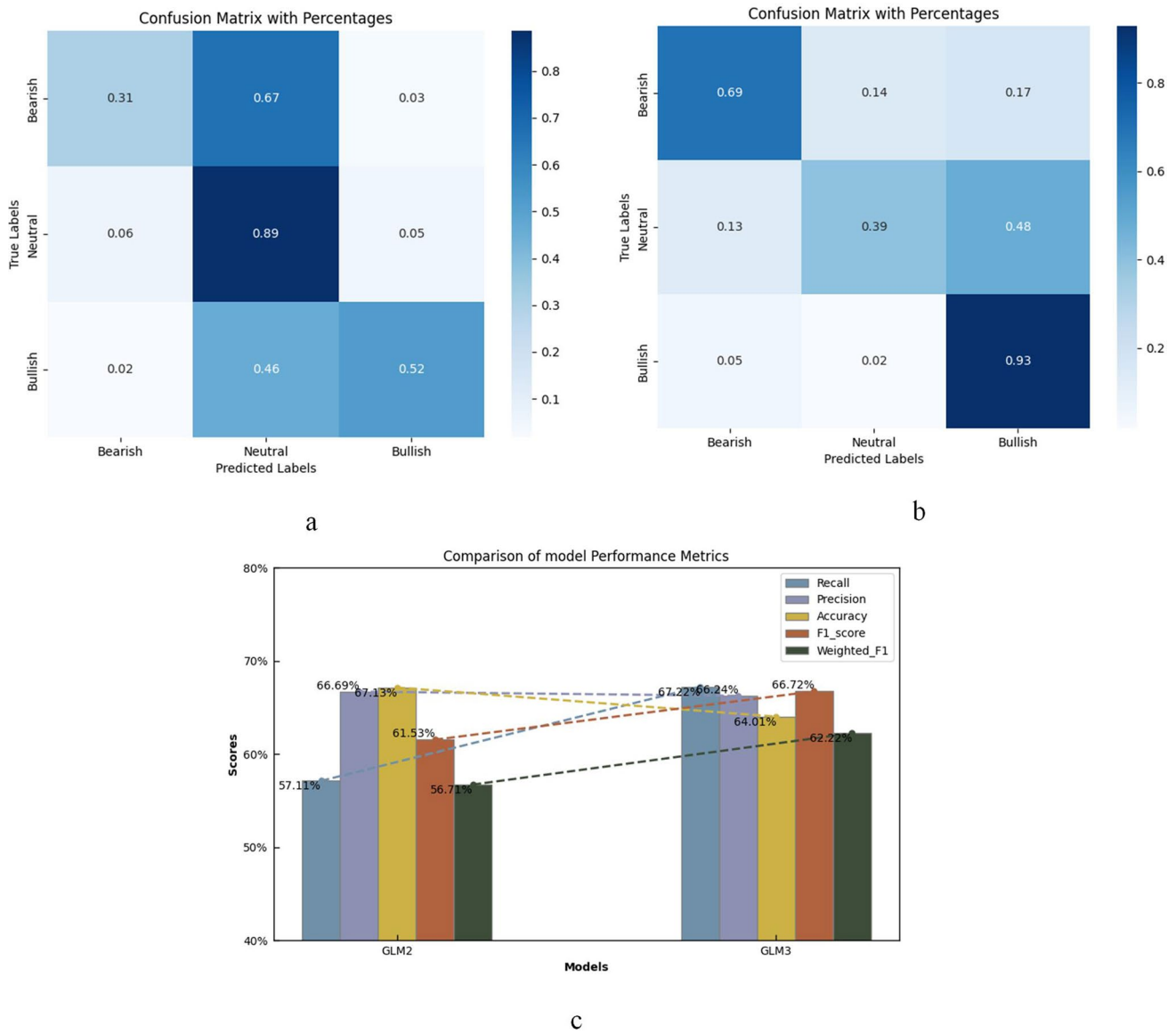


Fig. 13 Confusion matrix for CoT prompting on **a** GLM2, **b** GLM3, and **c** their performance comparison

approaches. Regarding prompt dependency, we conducted ablation studies across multiple prompting strategies, and consistently observed superior performance of DK-CoT, indicating relative robustness to prompt variation. Finally, to improve generalization, we evaluated our method across diverse financial news samples, including out-of-distribution texts, showing that DK-CoT maintains stable performance in realistic and dynamic financial scenarios. Based on the experimental results reported in Sect. 5.1, it can be revealed that the proposed DK-CoT prompting strategy can enhance the LLM's performance, especially in more advanced models, outperforming other prompting techniques in most cases. Some reflections are derived from the results:

- *The relationship between prompt engineering and LLMs* By using different prompt engineering techniques, the performance of LLMs will be improved to different degrees. Among them, the prompt templates designed by GLM2 and GLM3 under the techniques of few-shot prompting and DK-CoT prompts respectively achieve the best performance of the model. It suggests that adding the few-shot element in the prompt design would benefit the LLM's performance, and the insertion of domain knowledge would benefit more advanced LLMs. This may be because more advanced LLMs are trained with more quality data and equipped with the ability of a better understanding of external knowledge.

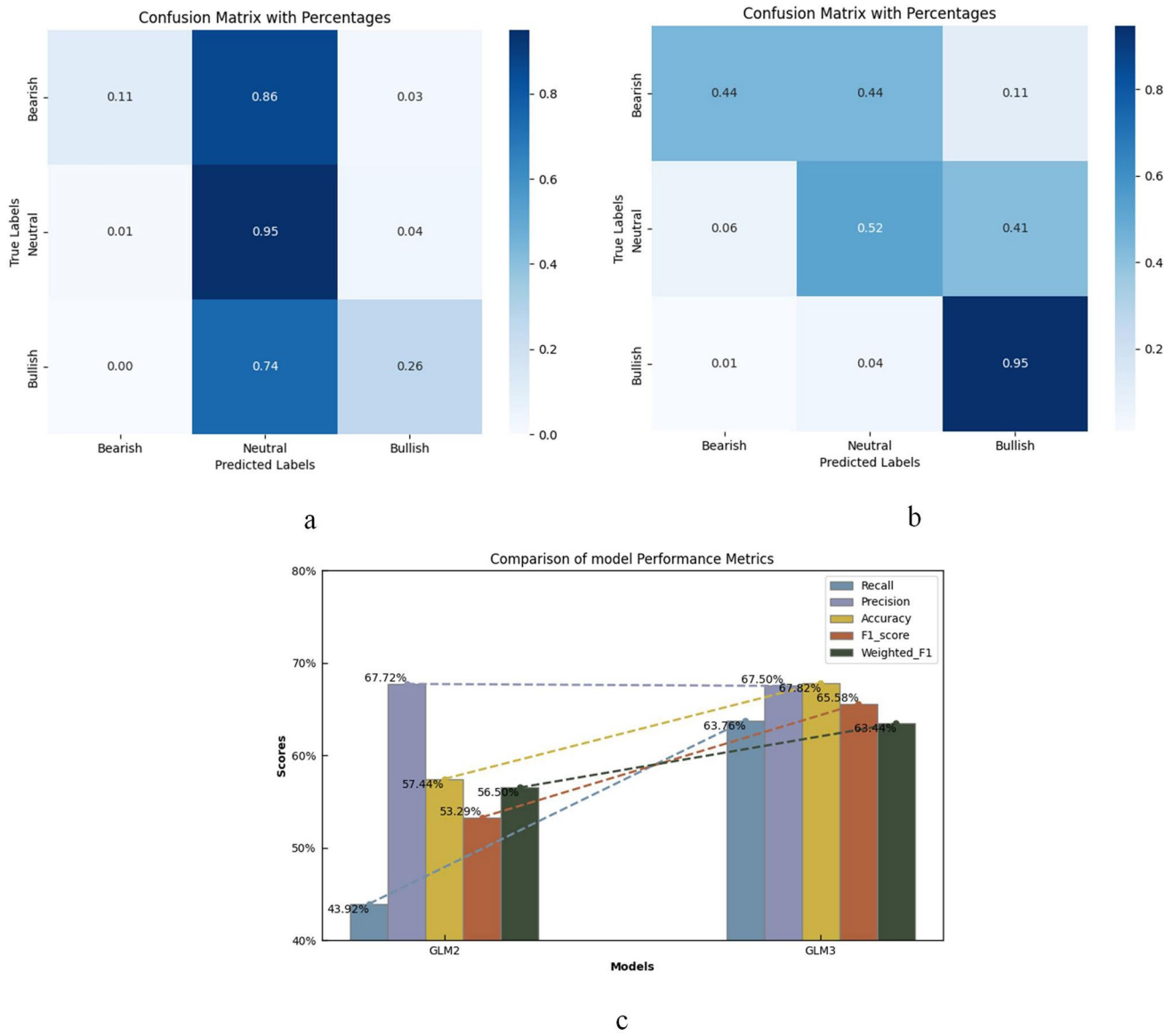


Fig. 14 Confusion matrix for generative knowledge prompting on **a** GLM2, **b** GLM3, and **c** their performance comparison

- *Ensemble prompting* Among numerous templates, GLM2 performs the best when using a prompting template designed with the few-shot prompting technique, while GLM3 performs the best with a prompting template designed using DK-CoT. Moreover, under the same prompting technique, the performance of GLM3 is superior to GLM2. Different prompt techniques have their own strengths. Using ensemble multiple prompt strategies may enhance the consistency of LLM performance.
- *Impact of model parameters and scale on Performance* As model parameters and scale increase, model performance generally improves. This phenomenon is validated in BERT, FinBERT, FLANG-RoBERTa, and Fin-FLANG-RoBERTa, GLM2, and GLM3 models, indicating that larger-scale models have higher potential and stability in handling financial news sentiment analysis tasks.
- *Compatibility of prompt techniques with LLM scale* More complex and advanced prompt techniques are more suitable for advanced LLMs, leading to more significant performance improvements. In contrast, basic prompt techniques are better suited for low-level LLMs. Applying advanced prompt techniques to lower-level LLMs may degrade performance while applying basic prompt techniques to advanced LLMs might not fully exploit the model's potential. Thus, for specific tasks, it is essential to choose suitable prompt techniques based on the LLM scale and design appropriate prompt templates to optimize model performance.

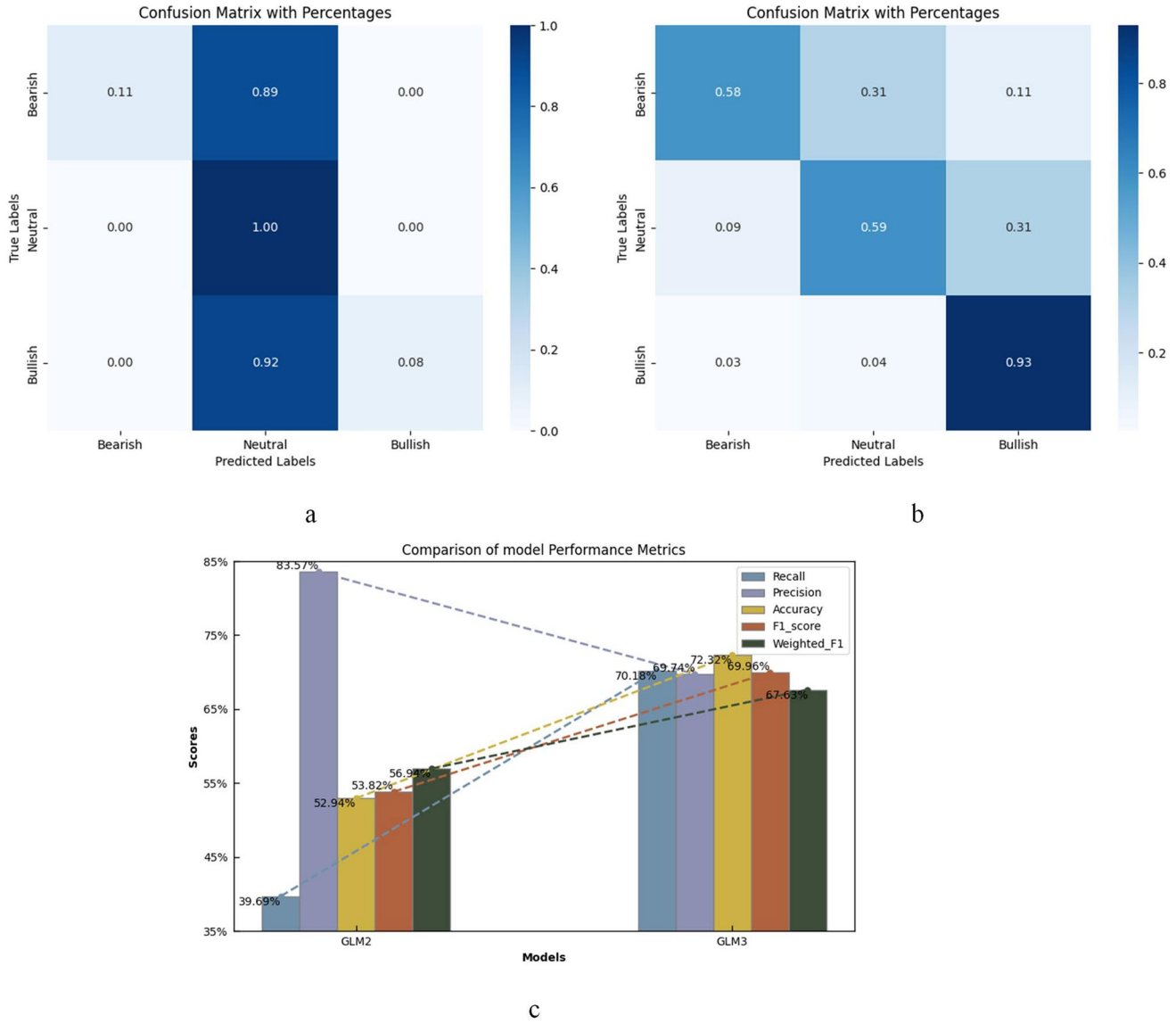


Fig. 15 Confusion matrix for DK-CoT prompting on **a** GLM2, **b** GLM3, and **c** their performance comparison

- *Weighted F1 metric* The proposal of the weighted F1 score as an evaluation metric for financial news sentiment analysis is justified by its ability to reflect the varying impacts of different sentiments on market behavior, a concept well-grounded in financial knowledge. Negative news tends to cause significant market reactions due to loss aversion, where the pain of losses is felt more acutely than the pleasure of gains, leading to stronger and often more immediate market movements. Conversely, positive news, while impactful, generally elicits more moderate responses, and neutral news has minimal influence. By assigning greater weight to negative sentiment, the weighted F1 score ensures that the evaluation metric aligns with the practical implications of sentiment misclassification in financial contexts, thus enhancing the reliability of the model for market predictions. This approach is supported by research in behavioral finance, highlighting the differential impact of news types on market volatility [35, 37]. Adopting the weighted F1 score thus ensures that the model's performance metrics are more representative of real-world financial decision-making processes.

On top of the comparative analysis in this study, we have also done some additional explorations on further advancing our proposed DK-CoT prompting strategy, namely ensemble prompting and a variant of DK-CoT by standardizing the output (i.e. JSONizing"). These explorations do not form a major part of the study but shed some light on other possibilities for enhancement.

5.2.2 Exploration 1: ensemble prompting

The idea of ensemble prompting is derived from the ensemble model in the deep learning paradigm, which has multiple models in parallel and jointly generate the output. Ensemble prompting combines the outputs of multiple prompts by majority voting or averaging (Fig. 16).

In this exploration exercise, we have integrated the two best prompting templates, namely few-shot and DK-CoT on both GLM2 and GLM3. Figure 17 depicts the experimental results. Although GLM's performance under the ensemble prompting strategy did not exceed the best-performing individual prompting template, it closely approached the performance of the best-performing prompting template. Moreover, under this strategy, both GLM2 and GLM3 outperformed the other prompting templates, indicating the effectiveness of the integrated prompting template strategy in enhancing LLM performance. It should be noted that this method can be easily extended to multiple prompts fed into various LLMs to jointly generate an "agreed" output, phasing out the effect of LLM choice.

5.2.3 Exploration 2: variant of DK-CoT by "JSONizing" the output

All previous experiments adopted a format where the model directly outputs the sentiment category of the news without providing explanatory notes. However, the model occasionally deviates from instructions, resulting in outputs as lengthy textual descriptions. This deviation not only affects classification accuracy but also increases the workload for manually processing data. To address this issue, we designed a variant of DK-CoT which standardizes the output format. This exploration exercise indicates in the prompt that a JSON format should be used to output sentiments. This method is named DK-CoT-JSON, which simplifies the post-processing of model outputs by providing clear, structured data. The JSON format includes the following objects: subject, event, impact, and sentiment. The subject in the JSON is fixed as China Petrochemical Corporation (Sinopec) in this experiment. The LLM extracts the events involving the subject from the news, the impact is analyzed using the chain of thought method in DK-CoT based on the extracted events, and the sentiment is determined as positive, negative, or neutral through the analysis of the impact.

As shown in Fig. 18. GLM2 again follows the pattern seen in many other strategies, predicting "neutral" most of the time, whereas GLM3 overall performs well. While the performance did not surpass the original DK-CoT, the output contains detailed reasoning, which enhances the interpretability of the model.

5.2.4 Consolidated analysis

In this study, BERT, FinBERT, FLANG-RoBERTa, Fin-FLANG-RoBERTa, GLM 2, and GLM 3, models were evaluated based on various metrics including weighted F1 scores, aiming to identify the best-performing model and prompting strategy.

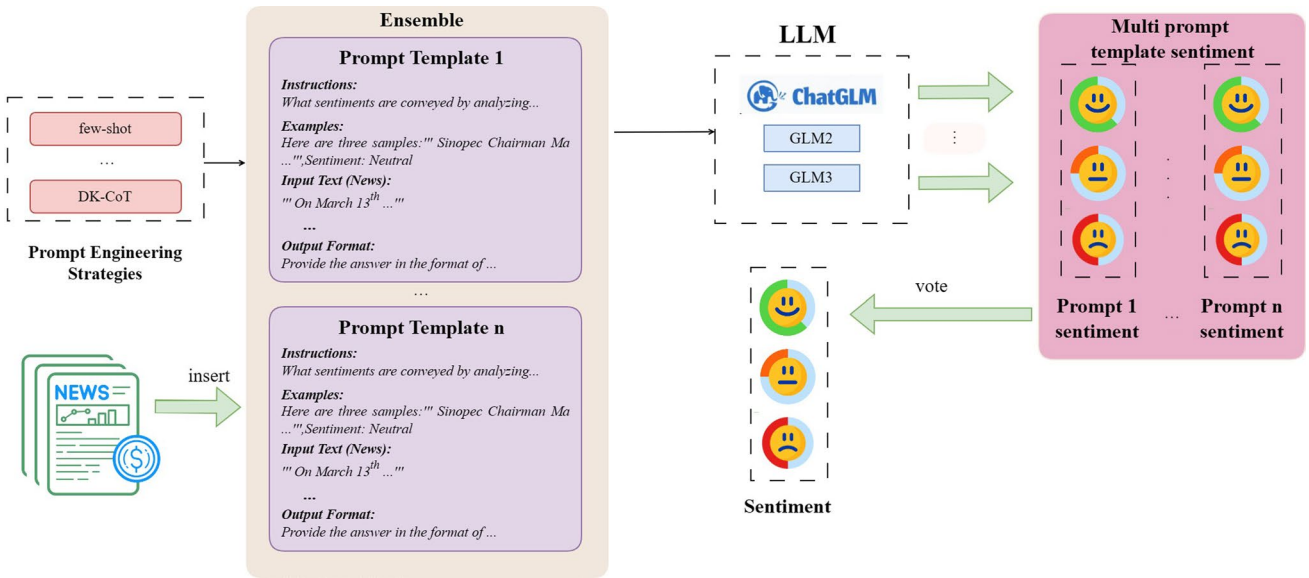


Fig. 16 Ensemble prompting flow chart

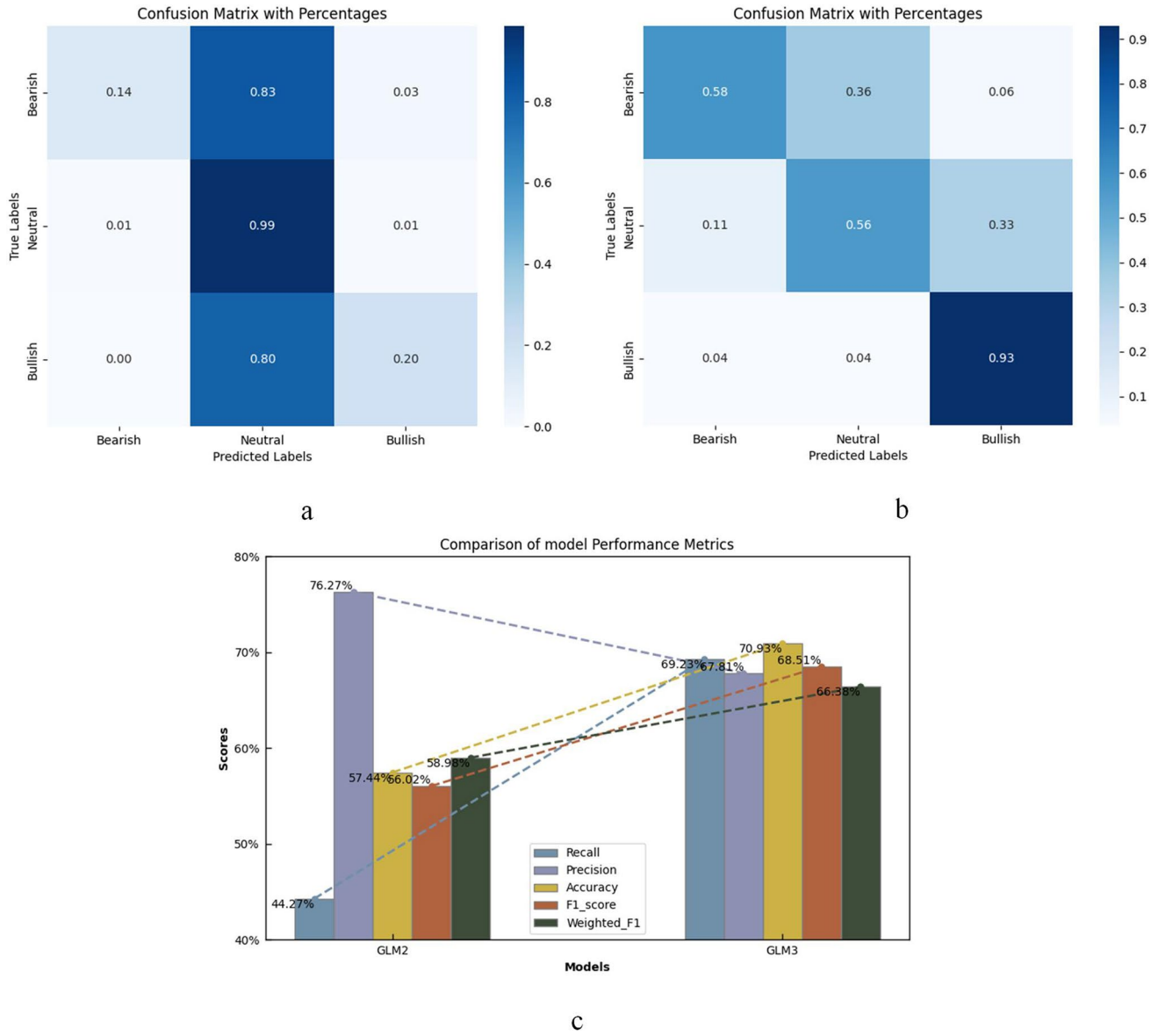


Fig. 17 Confusion matrix for ensemble prompting on **a** GLM2, **b** GLM3, and **c** their performance comparison

The weighted F1 scores for each model are presented in Fig. 19. Figure 20 and Table 2 comprehensively summarize the performance of all evaluated models and prompting strategies. GLM3 with the DK-CoT prompting performs best among all tested methods.

Through the experiments and explorations in this study, it is found that factors such as prompt engineering, integrated strategies, model scale, and fine-tuning effects significantly influence model performance. Different prompt techniques have different compatibilities with models of various scales, making reasonable selection and design of prompt templates crucial. Integrated strategies show great potential in enhancing LLM performance.

5.2.5 Significance test

An additional significance test has been conducted to rigorously evaluate the performance gains of our DK-CoT strategy compared to a baseline method (zero-shot). Specifically, we employed statistical tests such as paired

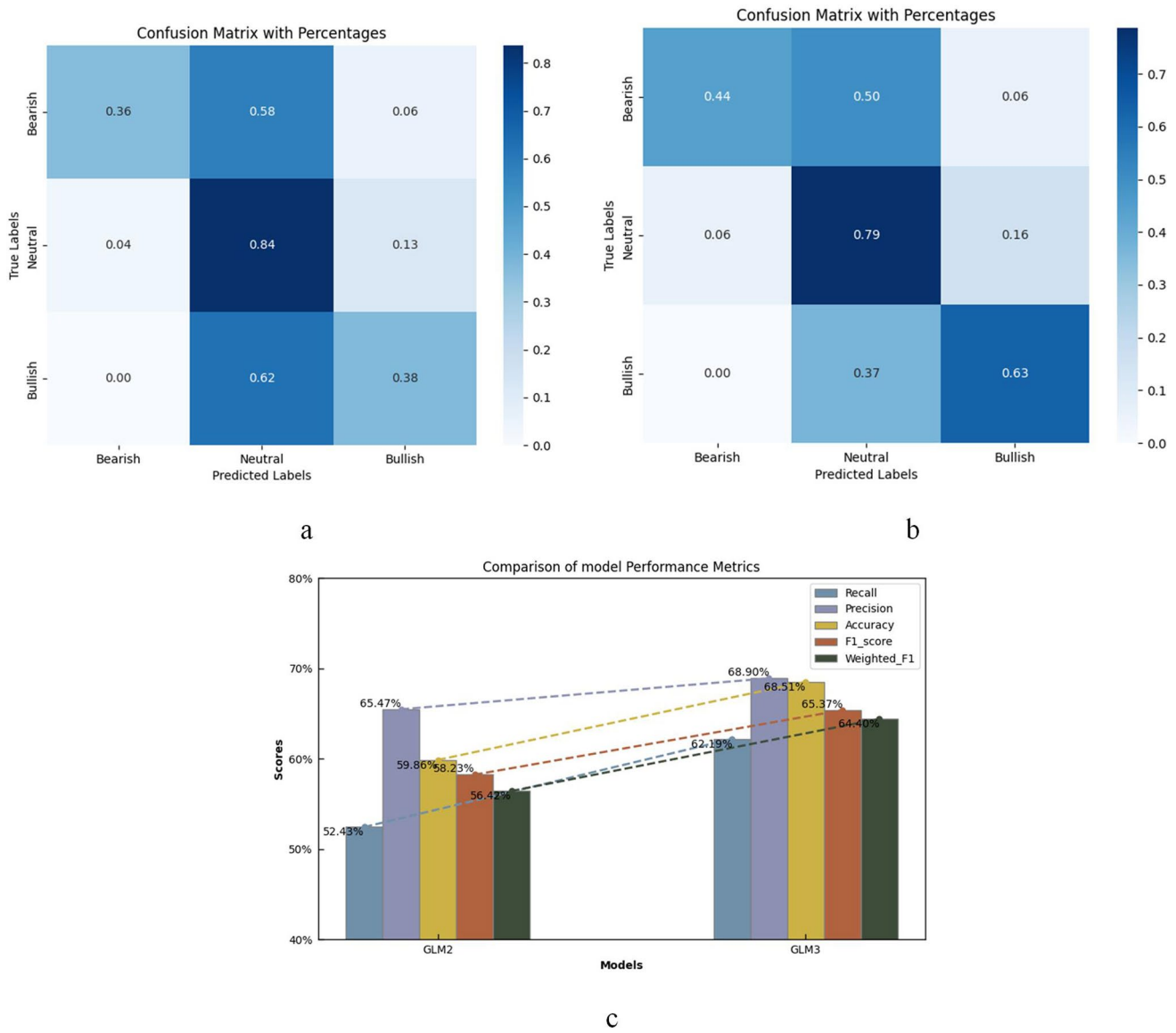


Fig. 18 Confusion matrix for DK-CoT-JSON prompting on **a** GLM2, **b** GLM3, and **c** their performance comparison

t-tests and Wilcoxon signed-rank tests to compare the performance metrics, i.e. weighted F1 scores, across multiple runs. The results of these tests confirm that the performance improvements achieved by the DK-CoT strategy are statistically significant. The p-values obtained from the tests are 4.18e-6 and 0.00195 respectively, which are both by far below the commonly accepted threshold ($p < 0.05$), indicating that the performance gains are not due to random chance but are a result of the enhanced prompt engineering approach.

5.2.6 Validation experiment

Through experiments on the Sinopec dataset, we found that the DK-CoT prompting technique maximally improved the performance of LLMs, surpassing models such as BERT, FinBERT, FLANG-RoBERTa, and Fin-FLANG-RoBERTa. To verify the effectiveness and stability of this prompting technique, further experiments were conducted using GLM2,

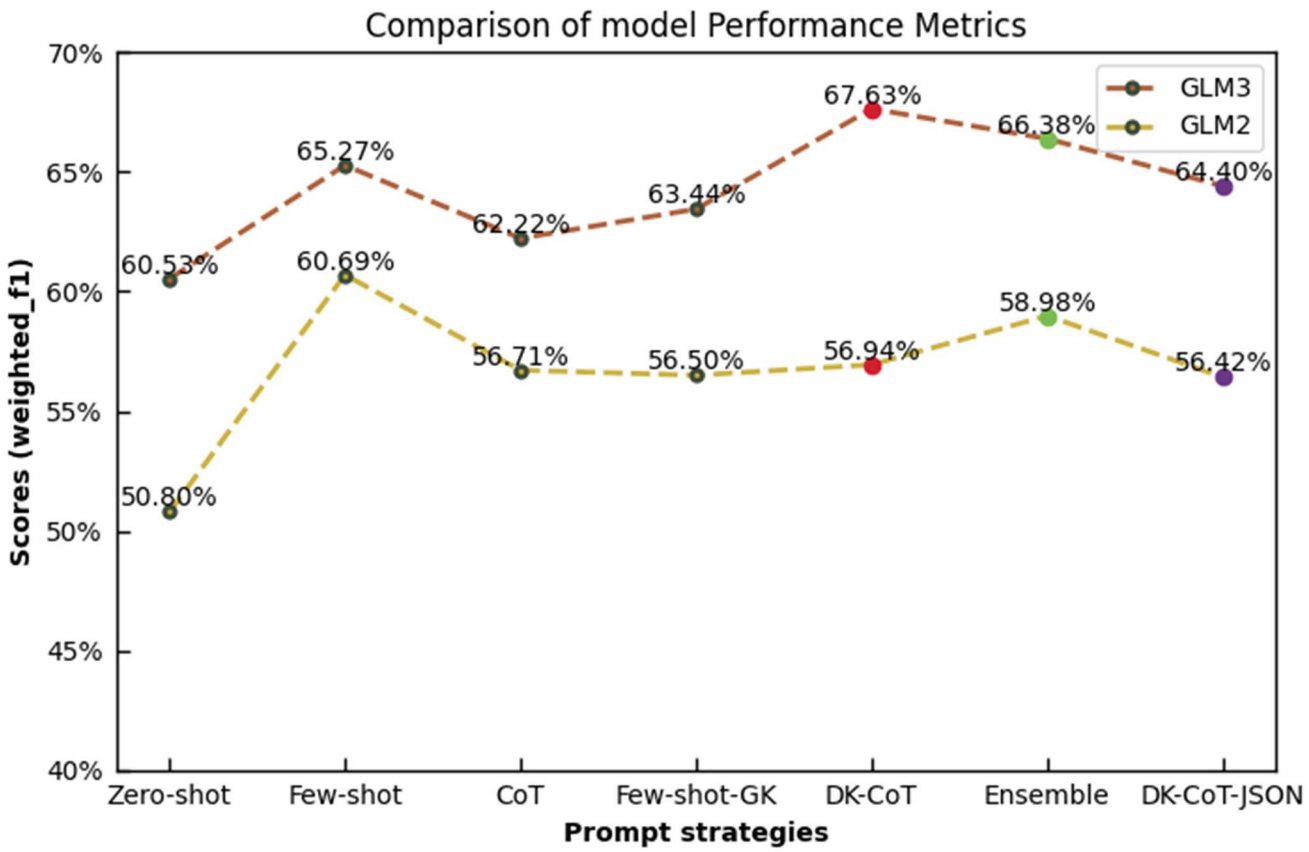


Fig. 19 Weighted F1 scores for various prompting strategies used on GLM2 and GLM3

GLM3, BERT, FinBERT, FLANG-RoBERTa, and Fin-FLANG-RoBERTa on three additional datasets: BYTTL, MI, and DI. The experimental results are shown in Table 3.

Figure 21 evaluates the performance of these models based on weighted F1 scores. The results indicate that GLM3, using the DK-CoT prompting technique, performed exceptionally well across different datasets. This further demonstrates the effectiveness and stability of the DK-CoT prompting technique in enhancing LLM performance on financial news sentiment analysis tasks, ensuring the reliability and comparability of the results.

6 Conclusion and future directions

The paper presents a novel prompt engineering strategy, DK-CoT, to enhance the performance of LLMs in financial news sentiment analysis, which has successfully addressed the research questions outlined in Sect. 1.2. For RQ 1, we explored the potential of LLMs such as GLM in financial news sentiment analysis, demonstrating that advanced models with extensive pre-training can be adapted via prompt engineering, domain knowledge insertion, and ensemble methods to provide precise and efficient sentiment predictions. The weighted F1 score has been proposed as a more practical evaluation metric, reflecting the greater impact of negative news on financial markets. This new metric ensures that the evaluation process aligns with real-world financial dynamics, providing a more accurate and practical assessment of sentiment analysis models. Our results show that LLMs can significantly enhance the accuracy of capturing market sentiments from financial news. For RQ 2, by comprehensive evaluation, DK-CoT improves the accuracy and reliability of sentiment analysis, outperforming benchmark models like BERT and RoBERTa, as well as other prompt engineering techniques such as zero-shot, few-shot, CoT, and generative knowledge prompting.

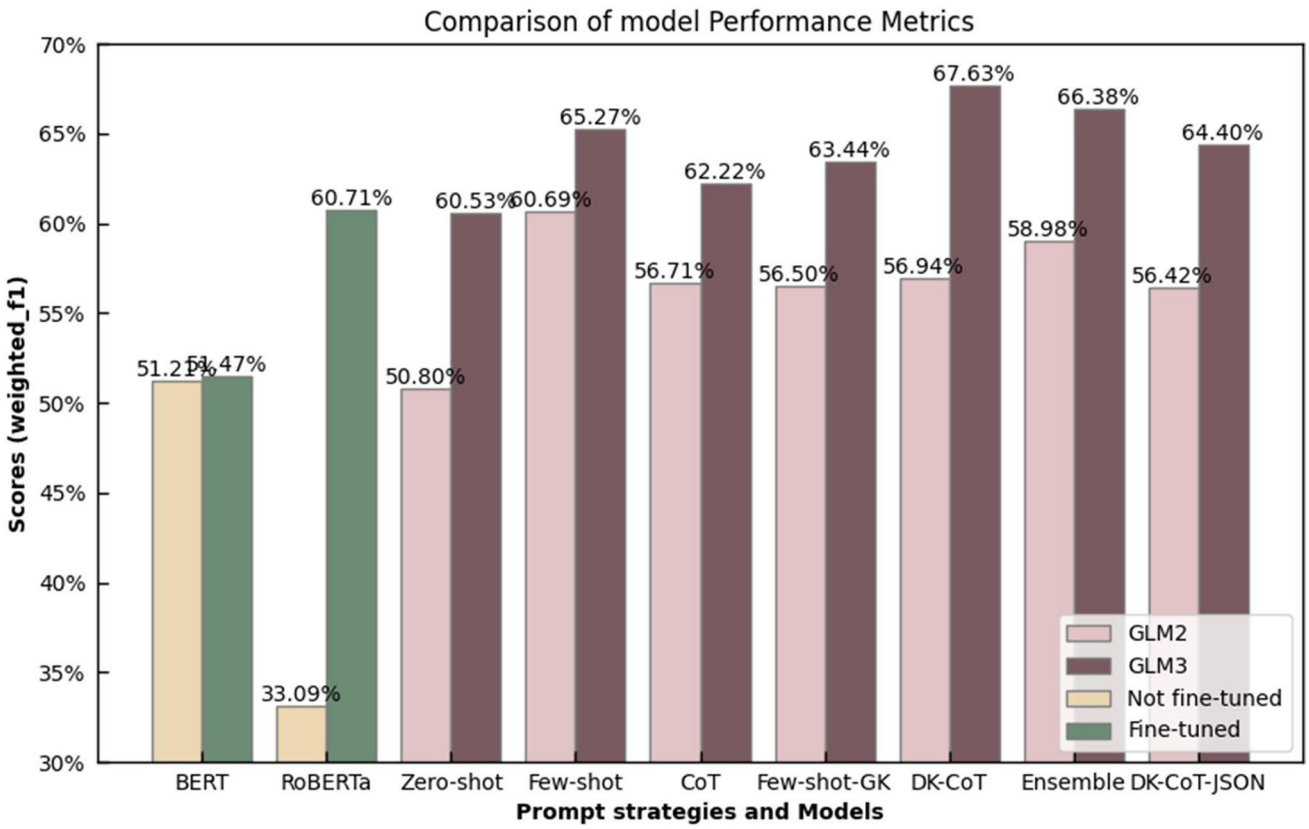


Fig. 20 The evaluation metrics for LLM and pre-trained language models (note: the fine-tuned BERT used in this study is FinBERT; the not-fine-tuned RoBERTa used in this study is the vanilla FLANG-RoBERTa, and the fine-tuned RoBERTa is Fin-FLANG-RoBERTa)

Table 2 Comparison of all tested models

Model	Method	Recall	Precision	Accuracy	F1	Weighted F1
BERT	Not fine-tuned	55.87%	58.75%	65.74%	56.63%	51.21%
FinBERT	Fine-tuned	76.11%	79.63%	72.32%	70.09%	51.47%
FLANG-RoBERTa	Not fine-tuned	7.08%	100.00%	51.21%	37.39%	33.09%
Fin-FLANG-RoBERTa	Fine-tuned	63.72%	77.42%	69.55%	68.52%	60.71%
GLM2	Zero-shot	48.21%	63.51%	59.86%	54.81%	50.80%
	Few-shot	57.52%	62.88%	65.05%	60.08%	60.69%
	CoT	57.11%	66.69%	67.13%	61.53%	56.71%
	Few-shot-GK	43.92%	67.72%	57.44%	53.29%	56.50%
	DK-CoT	39.69%	83.57%	52.94%	53.82%	56.94%
	Ensemble	44.27%	76.27%	57.44%	56.02%	58.98%
	DK-CoT variant	52.43%	65.47%	59.86%	58.23%	56.42%
GLM3	Zero-shot	61.00%	63.08%	60.90%	62.02%	60.53%
	Few-shot	67.87%	65.79%	69.20%	66.81%	65.27%
	CoT	67.22%	66.24%	64.01%	66.72%	62.22%
	Few-shot-GK	63.76%	67.50%	67.82%	65.58%	63.44%
	DK-CoT	70.18%	69.74%	72.32%	69.96%	67.63%
	Ensemble	69.23%	67.81%	70.93%	68.51%	66.38%
	DK-CoT variant	62.19%	68.90%	68.51%	65.37%	64.40%

Table 3 Evaluation metrics for revalidating datasets

Model	Method	Dataset	Recall	Precision	Accuracy	F1	Weighted F1
BERT	Not fine-tuned	MI	47.04%	51.33%	58.80%	45.91%	40.79%
		BYTTL	33.73%	33.28%	79.17%	32.59%	28.99%
		DI	51.42%	54.85%	63.44%	51.36%	45.93%
FinBERT	Fine-tuned	MI	38.74%	72.27%	63.34%	59.88%	51.19%
		BYTTL	43.75%	16.67%	56.67%	61.06%	27.48%
		DI	51.00%	66.54%	65.21%	63.01%	48.23%
FLANG-RoBERTa	Not fine-tuned	MI	4.05%	69.23%	51.54%	37.30%	29.73%
		BYTTL	0.00%	0.00%	82.50%	74.59%	27.12%
		DI	4.84%	80.95%	55.31%	41.49%	31.62%
Fin-FLANG-RoBERTa	Fine-tuned	MI	62.16%	63.89%	64.07%	63.42%	56.37%
		BYTTL	31.25%	83.33%	88.33%	86.19%	77.22%
		DI	61.25%	68.25%	68.75%	67.86%	59.97%
GLM2	DK-CoT	MI	53.54%	64.82%	53.54%	41.57%	46.89%
		BYTTL	85.83%	75.01%	85.83%	79.98%	59.35%
		DI	55.21%	66.53%	55.21%	41.11%	45.48%
GLM3	DK-CoT	MI	61.89%	66.53%	61.89%	60.56%	62.06%
		BYTTL	81.67%	88.58%	81.67%	83.69%	84.34%
		DI	64.48%	68.47%	64.48%	63.86%	62.60%

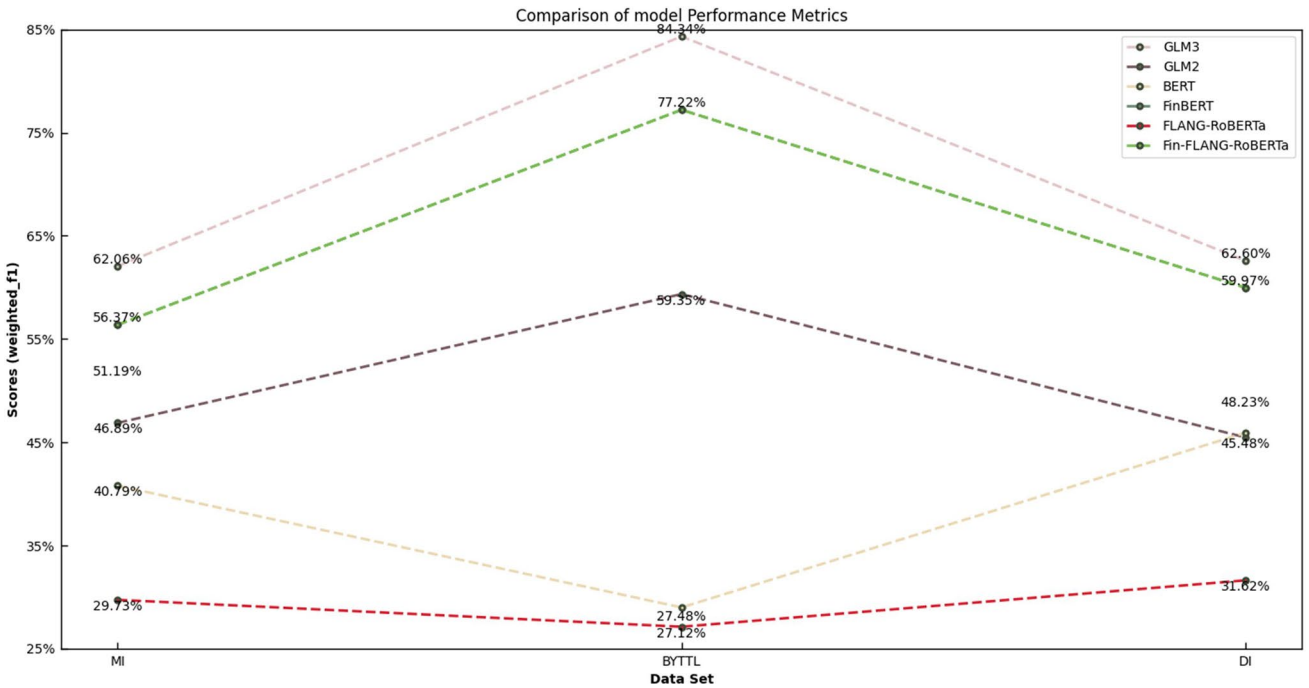


Fig. 21 The performance of LLM and pre-trained language models on different datasets

The study highlights the importance of incorporating domain-specific knowledge in prompt designs to improve LLM performance in predicting stock market movements. The DK-CoT strategy not only enhances the precision of sentiment analysis but also promotes sustainable AI practices by reducing the need for extensive computational resources and fine-tuning. Our results emphasize the practical applications of advanced prompt engineering techniques in financial decision-making and risk management, offering more accurate and timely insights into market sentiments.

The promising results of the DK-CoT strategy open several avenues for future research and development. First, future research can extend the DK-CoT strategy to other financial domains, such as earnings reports, social media sentiment, and analyst reports, to evaluate its generalizability and effectiveness across different types of financial texts. Also, other LLMs, especially open-source ones, can be employed for further comparison. Second, integrating the DK-CoT strategy with real-time financial news and market data may further enhance its applicability [38]. Developing systems that continuously update and adapt the domain knowledge component based on the latest financial news could provide even more timely and accurate sentiment predictions. In addition, the system can be extended further to dynamically retrieve domain knowledge from an established knowledge base or knowledge graph. Third, addressing ethical considerations and potential biases in financial sentiment analysis is crucial [39]. Future research can focus on identifying and mitigating biases in LLMs and prompt engineering strategies to ensure fair and unbiased sentiment predictions. Developing frameworks for ethical AI in financial analysis can enhance trust and reliability in these technologies.

Finally, we would like to highlight another promising direction worth thorough exploration, i.e. leveraging LLMs for forecasting tasks, like stock trend prediction [36]. There are two potential routes for this approach: merging sentiment data with fundamental financial metrics such as prices and trading volumes and feeding this combined data into downstream forecasting models; and developing systems where LLMs fetch sufficient real-time information, including news, reports, social media, and basic financial data, to provide predictions directly without the need for additional downstream models [40]. These approaches could harness the comprehensive understanding and real-time processing capabilities of LLMs, leading to more accurate and timely forecasts, thereby enhancing financial decision-making and market analysis.

Author contributions W.C., W.L., and X.Z. contributed to the study's conception and design. W.C., W.L., and J.Z. performed material preparation, data collection, and analysis. W.C. and W.L. wrote the first draft of the manuscript. W.C., W.L., and X.Z. commented on previous versions of the manuscript. W.C. and X.Z. provided the conceptualization, methodology, supervision, and funding. All authors read and approved the final manuscript.

Funding This research was supported by the Fujian Provincial Natural Science Foundation of China (Grant No. 2022J05291).

Data availability The data for this study was sourced from the Baidu Stock Market platform, which can be made available through the authors upon reasonable request.

Code availability The code for this study is available at https://github.com/XMUT-Service-Computing-Group/GLM-Sentiment.git.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Appendix

Prompt template for Sinopec.

Prompt (Chinese): 已知中国石油化工股份有限公司的产品主要包括原油、天然气、化纤、化肥、橡胶、成品油等。其竞争对手主要包括中国石油、中海油、中国化工等其他石油和化工企业，以及一些国际上的能源和化工巨头，如美国埃克森美孚、道达等。金融新闻对股票市场的影响是复杂而多样的。新闻信息会影响投资者情绪。积极新闻提升信心，推动股价上涨；消极新闻引发恐慌，导致股价下跌。中性新闻则被视为市场常态，投资者保持观望，股价平稳。这类新闻提供重要信息，但不会引起剧烈波动。融资消息对股市的影响取决于规模和用途。大规模买入且用途积极通常推动股价上涨，而小规模买入或用途不利则可能导致股价下跌。卖出消息若规模大且符合公司战略，通常也会推动股价上涨；若规模小或用途不明确，则可能引发股价下跌。适中规模的买入或卖出消息一般被视为中性，股价波动平稳。中国石化集团在 2023 年 11 月 11 日宣布计划在未来 12 个月内增持公司股份，金额在人民币 10 亿元至 20 亿元之间。截至 2024 年 9 月 11 日，已增持 273,371,456 股，占总股份的 0.22%，累计金额为人民币 1,187,146,632.57 元（不含税费）。此外，2023 年中国石化还将加大对新能源领域，特别是氢能的投资。这些举措符合国家能源结构转型政策，增强了公司的可持续发展能力和市场竞争力。中国政府提出“碳达峰、碳中和”目标，推动中国石化加快转型，减少碳排放并开发清洁能源。同时，政府重视能源安全，促进中国石化增加国内油气田开发，提升自主能源供应能力。在“十四五”期间，政府推动石化化工行业高质量发展，提出创新发展、产业结构优化、数字化转型和绿色安全等指导意见，以提升自主创新能力、提高产能利用率、推动数字化转型和绿色低碳发展。在融资活动方面，中国石化通过多种方式进行资本筹集，包括发行债券、股票增发和银团贷款等。中国石化实施 A+H 股份回购，截至 2023 年 10 月 26 日，公司回购 A 股已支付的总金额为 4.74 亿元，回购 H 股已支付的总金额为 2.88 亿港元（约合 2.7 亿元），两个月 A+H 股份回购金额达 7.44 亿元。中国石化的竞争对手中国石油也在加快能源转型的步伐，并在天然气、清洁能源开发等领域展开竞争。如中国石油在近年来加大了天然气领域的投资，提升其在国内能源市场的份额；加强对技术创新的投入，尤其是在提高采油效率和降低碳排放方面。中国石油还在研发更先进的油气勘探技术，以提升资源的开发能力。中海油则在在南海及其他海域进行深水钻探项目，提高公司的油气产量；在探索绿色能源领域的投资，特别是在海上风电项目方面。中海油计划通过在海上风电场的建设和运营，推进能源结构的优化。以下是三个样本：'"3 月 25 日，中国石化（600028）举行 2023 年度业绩说明会。中国石化董事长马永生回应中国石化 2023 年净利润出现下滑称，2023 年，国际原油价格总体呈震荡下行走势，同比下降 18.4%，境内化工市场供过于求、毛利仍处于低位。面对挑

战，公司全方位优化生产经营，大力协同攻坚创效，强化成本费用管控，推动全产业链提质增效，取得了来之不易的经营成果。马永生表示，公司净利润同比下降，主要是受库存变动减利以及计提矿业权出让收益等影响。若将以上因素剔除，同口径利润是大幅改善的。'"，原因分析：虽然新闻中提到了中国石化 2023 年净利润出现下滑，但同时也解释了下滑的原因，包括国际原油价格下降、境内化工市场供过于求等因素。此外，公司在面对挑战时采取了积极的措施，包括优化生产经营、强化成本费用管控等，取得了一定的经营成果。马永生还指出，如果剔除一些特定因素，同口径利润是大幅改善的，这暗示着公司在核心业务方面取得了进步。因此，这条新闻整体上可以被视为中性。情绪：中性；'"证券之星消息，中国石化 2023 年年报显示，公司主营收入 32122.15 亿元，同比下降 3.19%；归母净利润 604.63 亿元，同比下降 9.87%；扣非净利润 602.34 亿元，同比上升 3.92%；其中 2023 年第四季度，公司单季度主营收入 7422.74 亿元，同比下降 14.17%；单季度归母净利润 74.97 亿元，同比下降 23.79%；单季度扣非净利润 99.55 亿元，同比上升 417.14%；负债率 52.7%，投资收益 58.11 亿元，财务费用 99.22 亿元，毛利率 15.65%。'"，原因分析：主营收入和净利润都出现了同比下降，尤其是第四季度的数据显示了更为明显的下降趋势。尽管扣非净利润有所上升，但这可能受到特定因素的影响，而不代表整体业务状况的改善。负债率、财务费用等指标也需要关注，因为它们可能反映了公司的财务健康状况。综合来看，这则消息对中石化来说更多地倾向于坏消息。情绪：负面；'"3 月 13 日，中国石化与宁德时代新能源科技股份有限公司在北京签署战略合作框架协议。双方表示，将进一步深化战略合作，拓宽合作领域，延伸产业链条，加快转型升级步伐，推动双方合作迈上新台阶。'"，原因分析：这种增持表明南向资金对中国石油化工股份的信心增强，认为其具有投资价值。此外，股价上涨也反映了市场对公司业绩、前景或行业整体状况的乐观预期。港股通持股比例的增加也显示了市场对该公司的持续关注和投资意愿。因此，这一系列消息对中石化来说是好消息，它表明了市场对同行业公司的信心，也反映了整个行业的良好表现。情绪：正面。参考上述样本案例，结合上文提供的知识背景分析下文三引号中的新闻内容对中国石化传达了什么情绪？是正面、负面还是中性的？如果无法判断新闻内容表达的情绪，默认答案为中性。按照【情绪：正面】的格式给出答案，不要给出解释。'"新闻内容'"，情绪：

Prompt (English): China Petroleum Chemical Corporation (Sinopec) primarily produces crude oil, natural gas, synthetic fibers, fertilizers, rubber, and refined oil products. Its main competitors include China National Petroleum Corporation (CNPC), CNOOC, China National Chemical Corporation (ChemChina), and other petroleum and chemical enterprises, as well as international energy and chemical giants like ExxonMobil and Total. The impact of financial news on the stock market is complex and diverse. News information affects investor sentiment. Positive news boosts confidence and drives stock prices up; negative news triggers panic and causes stock prices to fall. Neutral news is viewed as the market norm, leading investors to adopt a wait-and-see approach, resulting in stable stock prices. Such news provides important information but does not cause dramatic fluctuations. The impact of financing news on the stock market depends on its scale and purpose. Large-scale buybacks with positive uses generally drive stock prices up, while small-scale buybacks or those with adverse uses may lead to a decline. Sell-off news, if large-scale and aligned with the company's strategy, usually drives stock prices up; if the scale is small or the purpose is unclear, it may cause stock prices to fall. Moderate-scale buy and sell news is generally considered neutral, with stock prices remaining stable. On November 11, 2023, Sinopec announced a plan to increase its shareholding in the company over the next 12 months, with an amount between RMB 1 billion and 2 billion. As of September 11, 2024, it has increased its holdings by 273,371,456 shares, accounting for 0.22% of the total shares, with a cumulative amount of RMB 1,187,146,632.57 (excluding taxes). Additionally, in 2023, Sinopec will increase its investment in the new energy sector, particularly in hydrogen energy. These measures align with the national energy structure transition policy and enhance the company's sustainability and market competitiveness. The Chinese government has set goals for "carbon peaking and carbon neutrality," pushing Sinopec to accelerate its transformation, reduce carbon emissions, and develop clean energy. At the same time, the government emphasizes energy security, promoting Sinopec to increase domestic oil and gas field development to enhance its energy supply capability. During the 14th Five-Year Plan period, the government is promoting high-quality development in the petrochemical industry, providing guidelines for innovative development, industrial structure optimization, digital transformation, and green safety to improve independent innovation capabilities, increase capacity utilization, and drive digital and green low-carbon development. In terms of financing activities, Sinopec raises capital through various methods, including bond issuance, stock increases, and syndicated loans. Sinopec has implemented an A+H share buyback program. As of October 26, 2023, the total amount spent on repurchasing A-shares was RMB 474 million, and the total amount spent on repurchasing H-shares was HKD 288 million (approximately RMB 270 million), with the total buyback amount for A+H shares reaching RMB 744 million over two months. Sinopec's competitor, CNPC, is also accelerating its energy transition efforts and competing in areas such as natural gas and clean energy development. For example, CNPC has increased its investment in the natural gas sector in

recent years to enhance its market share in the domestic energy market and strengthened its investment in technological innovation, especially in improving oil extraction efficiency and reducing carbon emissions. CNPC is also developing more advanced oil and gas exploration technologies to improve resource development capabilities. CNOOC is engaged in deep-water drilling projects in the South China Sea and other areas to increase its oil and gas production. It is also investing in green energy, particularly offshore wind power projects, with plans to optimize the energy structure through the construction and operation of offshore wind farms. Here are three samples: '"On March 25, Sinopec (600028) held the 2023 performance briefing. Sinopec Chairman Ma Yongsheng responded to the decline in Sinopec's 2023 net profit, stating that in 2023, international crude oil prices showed an overall downward trend, down 18.4% year-on-year, and the domestic chemical market was oversupplied, with gross margins remaining low. In the face of challenges, the company optimized production and operations across the board, worked collaboratively to achieve results, strengthened cost control, and promoted the entire industry chain to improve quality and efficiency, achieving hard-won operational results. Ma Yongsheng said that the year-on-year decline in net profit was mainly due to reduced profits from inventory changes and the accrual of mining rights transfer income. If these factors were excluded, the comparable profit would have improved significantly."' Reasoning: Although the news mentioned Sinopec's net profit decline in 2023, it also explained the reasons for the decline, including the drop in international crude oil prices and oversupply in the domestic chemical market. In addition, the company took proactive measures to optimize production and operations, strengthen cost control, and achieve certain results. Ma Yongsheng also pointed out that if specific factors were excluded, comparable profits would have improved significantly, implying that the company made progress in its core business. Therefore, this news can overall be considered neutral. Sentiment: Neutral; '"According to Securities Star, Sinopec's 2023 annual report shows that the company's main revenue was 3.212215 trillion yuan, a year-on-year decrease of 3.19%; net profit attributable to the parent was 60.463 billion yuan, a year-on-year decrease of 9.87%; non-recurring net profit was 60.234 billion yuan, a year-on-year increase of 3.92%; in the fourth quarter of 2023, the company's single-quarter main revenue was 742.274 billion yuan, a year-on-year decrease of 14.17%; single-quarter net profit attributable to the parent was 7.497 billion yuan, a year-on-year decrease of 23.79%; single-quarter non-recurring net profit was 9.955 billion yuan, a year-on-year increase of 417.14%; debt ratio was 52.7%, investment income was 5.811 billion yuan, financial expenses were 9.922 billion yuan, and gross margin was 15.65%."' Reasoning: Both the main revenue and net profit showed year-on-year declines, with a more pronounced downward trend in the fourth quarter. Although non-recurring net profit increased, this may have been influenced by specific factors and does not reflect an improvement in the overall business situation. Indicators such as the debt ratio and financial expenses also need attention, as they may reflect the company's financial health. Overall, this news leans more towards negative for Sinopec. Sentiment: Negative; '"On March 13, Sinopec signed a strategic cooperation framework agreement with Contemporary Amperex Technology Co., Ltd. (CATL) in Beijing. Both parties stated that they would further deepen strategic cooperation, broaden cooperation areas, extend the industrial chain, accelerate the pace of transformation and upgrading, and push their cooperation to a new level."' Reasoning: This increase in holdings shows that southbound capital's confidence in Sinopec has strengthened, believing that it has investment value. Additionally, the rise in stock prices reflects market optimism about the company's performance, prospects, or the overall industry situation. The increase in the proportion of Hong Kong Stock Connect holdings also shows the market's continuous attention and willingness to invest in the company. Therefore, this series of news is positive for Sinopec, as it indicates market confidence in its industry peers and reflects good performance across the industry. Sentiment: Positive. Based on the sample cases above, analyze what sentiment the following news conveys about Sinopec according to the provided background knowledge. Is it positive, negative, or neutral? If it is impossible to determine the sentiments expressed in the news content, the default answer is neutral. Provide the answer in the format of [Sentiment: Positive], without providing an explanation. '"News Full Text"', Sentiment:

Prompt template for Xiaomi Group.

Prompt (Chinese): 已知小米集团是专注于智能硬件和电子产品研发的全球化移动互联网企业，同时也是一家专注于智能手机、智能电动汽车、互联网电视及智能家居生态链建设的创新型科技企业。其主要竞争对手有华为、oppo、苹果和vivo 等公司，这些公司在智能手机、智能家居、智能穿戴设备等多个领域与小米集团竞争。新闻信息会影响投资者情绪。积极新闻提升信心，推动股价上涨；消极新闻引发恐慌，导致股价下跌。中性新闻则被视为市场常态，投资者保持观望，股价平稳。这类新闻提供重要信息，但不会引起剧烈波动。融资消息对股市的影响取决于规模和用途。大规模买入且用途积极通常推动股价上涨，而小规模买入或用途不利则可能导致股价下跌。卖出消息若规模大且符合公司战略，通常也会推动股价上涨；若规模小或用途不明确，则可能引发股价下跌。适中规模的买入或卖出消息一般被视为中性，股价波动平稳。小米集团受益于“一带一路”政策，其国际化收入在2018年第三季度同比增长112.7%，海外收入占比达到 43.9%，该政策对公司业务产生巨大帮助。国家发改委对小米的发展给予了积极评价，认为小米发展全生态链模式，推动了先进制造业和现代服务业的融合，这与国家推动两业融合发展的政策相契合。小米通过技术创新和自主研发，推动了智能终端和品质化发展，提升了“产品+服务”的融合发展水平，拓展了产业发展空间。小米集团在 2022 年 3 月 22 日宣布了一项股份回购计划，计划以最高总额 100 亿港元在公开市场购回股份。在 2024 年 3 月 28 日，小米汽车召开发布会，正式发布了小米 SU7 系列车型。小米 SU7 系列车型在设计、性能和续航方面都有显著特点，其中 Max 版续航达到810公里，搭载宁德时代麒麟电池，是国内唯一同时实现2秒级零百加速和超 800 公里续航的纯电动汽车。小米集团在2023 年经历了一系列高管调整和组织架构变动。2023 年 1月，小米集团晋升卢伟冰为集团总裁，同时晋升王晓雁、屈恒和马骥为集团副总裁。此外，小米集团合伙人、高级副总裁、大家电部总裁张峰于2023年12月完成工作交接后离职。小米集团的竞争对手苹果定期发布新款 iPhone，新产品的发布吸引大量消费者关注，对小米的市场份额构成挑战；一加手机在发布会上全面“对标”小米，这种直接的市场竞争行为影响小米的市场策略和消费者选择；vivo 官宣自研操作系统，在 2023 年开发者大会上正式发布了自研的蓝河操作系统，它引入蓝心大模型的能力，支持复杂的意图识别和声音、图片、手势等自由交互方式，并为开发者提供了自动编码等应用开发新范式，这直接影响小米在软件生态方面的策略。以下是三个样本：'"7月19日小米集团-WR发布公告称，公司于 2024 年 7 月 19 日在香港交易所回购 300.00 万股，耗资4963.96万港币，根据此次回购数量和耗资情况计算回购均价约为 16.55 港币；根据披露此次最高回购价 16.58 港币，最低回购价 16.52 港币。据了解，小米集团-WR 近三个月累计回购股份数为 9041.38 万股，占公司已发行股本的 0.36%。'"，原因分析：小米集团-WR 回购 300.00 万股，耗资 4963.96万港币。虽然回购行为通常被视为公司对自身股票价值的信

心表现，但回购数量相对于公司已发行股本的比例（0.36%）并不算大，因此对市场整体影响有限。且回购均价约为 16.55 港币，最高回购价为 16.58 港币，最低回购价为 16.52 港币。价格波动幅度较小，显示市场对该股票的定价较为稳定。虽然小米集团-WR 的回购行为表明了公司对自身股价的信心，但由于回购数量和比例较小，对市场的实际影响有限。情绪：中性；"'07 月 23 日,小米集团-W 股价跌 1.51%，报收 16.96 元，成交金额 10.99 亿元，换手率 0.32%，振幅 3.37%，量比 0.81。小米集团-W 今日主力资金（超大单+大单）净流出 9502 万元，上一交易日主力净流入 2.43 亿元。该股近 5 个交易日上涨 3.18%，主力资金累计净流入 2.27 亿元；近 20 日主力资金累计净流入 7.41 亿元，其中净流入天数为 14 日。'"，原因分析：股价下跌 1.51%，报收 16.96 元。股价下跌通常反映市场对公司短期前景的负面情绪。成交金额为 10.99 亿元，换手率为 0.32%，振幅为 3.37%，量比为 0.81。尽管成交量和换手率显示出市场活跃度，但股价下跌仍表明市场情绪不佳。尽管长期来看主力资金有累计净流入的趋势，但当天的股价下跌和主力资金的净流出是对市场情绪的负面影响。情绪：负面； "'7 月 29 日，小米集团-W(01810)盘中上涨 2.08%，截至 10:24，报 16.7 元/股，成交 3.01 亿元。小米集团是一家主要以手机、智能硬件和 IoT 平台为核心的互联网公司，产品包括智能手机、IoT 和生活消费产品、互联网服务产品。公司的商业模式由创新、高质量的硬件，高效新零售和丰富的互联网服务三个相互协作的支柱组成，致力于降低运营成本并提升效率。截至 2024 年一季报，小米集团-W 营业总收入 755.07 亿元、净利润 41.82 亿元。'"，原因分析：小米集团-W 盘中上涨 2.08%，报 16.7 元/股。股价的上涨通常反映出市场对公司的乐观情绪。且成交金额达到 3.01 亿元，显示出较高的市场活跃度和投资者的关注。股价的上涨、较高的成交额以及公司的稳健业务模式和良好的财务表现都显示出市场和投资者对小米集团的积极态度和信心。情绪：正面。参考上述样本案例，结合上文提供的知识背景分析下文三引号中的新闻内容对中国石化传达了什么情绪？是正面、负面还是中性的？如果无法判断新闻内容表达的情绪，默认答案为中性。按照【情绪：正面】的格式给出答案，不要给出解释。"'新闻内容'"，情绪：

Prompt (English): Xiaomi Group is a global mobile internet enterprise focused on the research and development of smart hardware and electronic products. It is also an innovative technology company specializing in smartphones, smart electric vehicles, internet televisions, and smart home ecosystem construction. Major competitors include Huawei, Oppo, Apple, and Vivo, which compete with Xiaomi in various fields such as smartphones, smart homes, and wearable devices. News information can impact investor sentiment. Positive news boosts confidence and drives stock prices up; negative news triggers panic and causes stock prices to fall. Neutral news is seen as the market norm, leading investors to adopt a wait-and-see approach, resulting in stable stock prices. Such news provides important information but does not cause dramatic fluctuations. Financing news affects the stock market based on its scale and purpose. Large-scale buybacks with positive uses typically drive stock prices up, while small-scale buybacks or those with adverse uses may lead to a decline. Sell-off news, if large-scale and aligned with the company's strategy, usually drives stock prices up; if the scale is small or the purpose is unclear, it may cause stock prices to fall. Moderate-scale buy and sell news is generally considered neutral, with stock prices remaining stable. Xiaomi Group benefits from the "Belt and Road" policy, with its international revenue increasing by 112.7% year-on-year in the third quarter of 2018, and overseas revenue accounting for 43.9% of total revenue. This policy has significantly aided the company's business. The National Development and Reform Commission has given Xiaomi a positive evaluation, noting that Xiaomi's development of a full ecosystem model promotes the integration of advanced manufacturing and modern services, aligning with the national policy of integrating these two sectors. Through technological innovation and independent research and development, Xiaomi has advanced smart terminals and quality development, enhancing the integration of "products + services" and expanding industrial development space. On March 22, 2022, Xiaomi Group announced a share buyback plan, intending to repurchase shares up to a maximum total of 10 billion HKD in the open market. On March 28, 2024, Xiaomi Auto held a press conference to officially launch the Xiaomi SU7 series models. The Xiaomi SU7 series stands out in design, performance, and range, with the Max version achieving a range of 810 kilometers, equipped with CATL's Kirin battery. It is the only domestic pure electric vehicle to simultaneously achieve 2-second acceleration and over 800 kilometers of range. In 2023, Xiaomi Group underwent a series of executive adjustments and organizational changes. In January 2023, Lu Weibing was promoted to Group President, while Wang Xiaoyan, Qu Heng, and Ma Ji were promoted to Group Vice Presidents. Additionally, Xiaomi Group partner and Senior Vice President Zhang Feng, who was also President of the Large Home Appliances Division, left the company after completing his work handover in December 2023. Xiaomi's competitor, Apple, regularly releases new iPhone models, attracting significant consumer attention and challenging Xiaomi's market share. OnePlus has fully "targeted" Xiaomi in its launch events, impacting Xiaomi's market strategy and consumer choices. Vivo announced its self-developed operating system, officially releasing the Blue River OS at the 2023 Developer Conference. This OS introduces the Blue Core model's capabilities, supports complex intent recognition and free interactions through voice, images, and gestures, and offers new paradigms for application development, directly influencing Xiaomi's software ecosystem strategy. Here are three examples: "'On July 19, Xiaomi

Group-WR announced that the company repurchased 3 million shares on July 19, 2024, at the Hong Kong Stock Exchange, costing 49.64 million HKD. Based on the repurchase volume and cost, the average repurchase price was approximately 16.55 HKD, with the highest repurchase price being 16.58 HKD and the lowest at 16.52 HKD. It is understood that over the past three months, Xiaomi Group-WR has repurchased a total of 90.41 million shares, accounting for 0.36% of the company's issued capital.'" Reason Analysis: Xiaomi Group-WR repurchased 3 million shares at a cost of 49.64 million HKD. Although buybacks are usually seen as a sign of the company's confidence in its own stock value, the repurchased volume relative to the company's issued capital (0.36%) is not significant, so its overall market impact is limited. The average repurchase price was around 16.55 HKD, with a small price fluctuation between 16.52 and 16.58 HKD, indicating a stable market valuation for the stock. While the repurchase signals Xiaomi's confidence in its stock, the relatively small scale and proportion limit its actual impact on the market. Sentiment: Neutral. '"On July 23, Xiaomi Group-W's stock price dropped by 1.51%, closing at 16.96 RMB, with a trading volume of 1.099 billion RMB, a turnover rate of 0.32%, an amplitude of 3.37%, and a volume ratio of 0.81. The net outflow of main funds (super-large and large orders) was 95.02 million RMB, compared to the previous trading day's net inflow of 243 million RMB. Over the past 5 trading days, the stock has risen by 3.18%, with cumulative main fund net inflows of 227 million RMB. In the last 20 trading days, the cumulative main fund net inflow was 741 million RMB, with 14 days showing net inflows.'" Reason Analysis: The stock price dropped by 1.51%, closing at 16.96 RMB. Stock price declines generally reflect negative market sentiment toward the company's short-term outlook. The trading volume was 1.099 billion RMB, with a turnover rate of 0.32%, an amplitude of 3.37%, and a volume ratio of 0.81. Despite active trading, the price drop shows weak market sentiment. Although there has been a trend of net inflows in main funds over the long term, the stock's decline and net fund outflow on that day reflect negative market sentiment. Sentiment: Negative. '"On July 29, Xiaomi Group-W (01810) rose 2.08% during the session. As of 10:24, the stock was trading at 16.7 RMB per share, with a trading volume of 301 million RMB. Xiaomi Group is primarily an internet company centered on smartphones, smart hardware, and IoT platforms, offering products including smartphones, IoT, consumer electronics, and internet services. The company's business model is built on three pillars: innovative, high-quality hardware, efficient new retail, and rich internet services, aimed at reducing operational costs and increasing efficiency. As of its Q1 2024 report, Xiaomi Group-W had total revenue of 75.507 billion RMB and a net profit of 4.182 billion RMB.'" Reason Analysis: Xiaomi Group-W's stock price rose by 2.08%, trading at 16.7 RMB per share. The rise in stock price typically reflects positive market sentiment toward the company. The trading volume of 301 million RMB indicates high market activity and investor interest. The rise in stock price, significant trading volume, and the company's strong business model and financial performance demonstrate the market's and investors' positive attitude and confidence in Xiaomi Group. Sentiment: Positive. Based on the sample cases above, analyze what sentiment the following news conveys about Sinopec according to the provided background knowledge. Is it positive, negative, or neutral? If it is impossible to determine the sentiments expressed in the news content, the default answer is neutral. Provide the answer in the format of [Sentiment: Positive], without providing an explanation. '"News Full Text"', Sentiment:

References

1. Cui J, Wang Z, Ho S-B, Cambria E. Survey on sentiment analysis: evolution of research methods and topics. Artif Intell Rev. 2023;56(8):8469–510. https://doi.org/10.1007/s10462-022-10386-z.
2. Chen W, Hussain W, Cauteruccio F, Zhang X. Deep learning for financial time series prediction: a state-of-the-art review of standalone and hybrid models. Comput Model Eng Sci. 2024;139(1):187–224.
3. G. Team *et al.*, "ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools," *arXiv e-prints,* p. arXiv: 2406.12793, 2024.
4. W. Chen, I. Al-Qudah, and F. Rabhi, "A Framework for Facilitating Reproducible News Sentiment Impact Analysis," presented at the Proceedings of the 2022 5th International Conference on Software Engineering and Information Management, Yokohama, Japan, 2022. https://doi.org/10.1145/3520084.3520104.
5. Yazdani SF, Murad MAA, Sharef NM, Singh YP, Latiff ARA. Sentiment classification of financial news using statistical features. Int J Pattern Recognit Artif Intell. 2017;31(03):1750006. https://doi.org/10.1142/s0218001417500069.
6. Ahmad HO, Umar SU. Sentiment analysis of financial textual data using machine learning and deep learning models. Informatica. 2023. https://doi.org/10.31449/inf.v47i5.4673.
7. Das N, Sadhukhan B, Bhakta SS, Chakrabarti S. Integrating EEMD and ensemble CNN with X (Twitter) sentiment for enhanced stock price predictions. Soc Network Anal Mining. 2024. https://doi.org/10.1007/s13278-023-01190-w.
8. Li W, Zhu L, Shi Y, Guo K, Cambria E. User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models. Appl Soft Comput. 2020. https://doi.org/10.1016/j.asoc.2020.106435.
9. Sharaff A, Chowdhury TR, Bhandarkar S. LSTM based sentiment analysis of financial news. SN Comput Sci. 2023. https://doi.org/10.1007/s42979-023-02018-2.
10. Lengkeek M, van der Knaap F, Frasincar F. Leveraging hierarchical language models for aspect-based sentiment analysis on financial data. Inf Proc Manag. 2023. https://doi.org/10.1016/j.ipm.2023.103435.
11. Leippold M. Sentiment spin: attacking financial sentiment with GPT-3. Finance Res Lett. 2023. https://doi.org/10.1016/j.frl.2023.103957.
12. Zhang B, Yang H, Zhou T, Babar MA, Liu X-Y. Enhancing financial sentiment analysis via retrieval augmented large language models. Brooklyn: NY, USA; 2023.
13. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng. 2023. https://doi.org/10.1007/s10439-023-03272-4.
14. Fatouros G, Soldatos J, Kouroumali K, Makridis G, Kyriazis D. Transforming sentiment analysis in the financial domain with ChatGPT. Machine Learn Applic. 2023. https://doi.org/10.1016/j.mlwa.2023.100508.
15. H. Zhang, F. Hua, C. Xu, J. Guo, H. Kong, and R. Zuo, "Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements?," *arXiv preprint* arXiv:2306.14222*,* 2023.
16. Cao Y, Zhai J. Bridging the gap–the impact of ChatGPT on financial research. J Chinese Econ Bus Stud. 2023;21(2):177.
17. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55(9):1–35.
18. Ouyang L, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44.
19. D. Cao *et al.*, "Tempo: Prompt-based generative pre-trained transformer for time series forecasting," *arXiv preprint* arXiv:2310.04948*,* 2023.
20. B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *arXiv preprint* arXiv:2310.14735*,* 2023.
21. Yong G, Jeon K, Gil D, Lee G. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Comput-Aided Civil Infrast Eng. 2023;38(11):1536–54.
22. Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, Finbert: A pre-trained financial language representation model for financial text mining, in Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence. 2021.
23. Peng B, Chersoni E, Hsu Y-Y, Qiu L, Huang C-R. Supervised cross-momentum contrast: aligning representations with prototypical examples to enhance financial sentiment analysis. Knowl-Based Syst. 2024;295: 111683.
24. Adelakun N, Baale A. Sentiment analysis of financial news using the bert model. ITEGAM J Eng Technol Indust Applic. 2024. https://doi.org/10.5935/jetia.v10i48.1029.
25. Sinha A, Kedas S, Kumar R, Malo P. SEntFiN 1.0: entity-aware sentiment analysis for financial news. J Am Soc Inf Sci. 2022;73(9):1314–35.
26. K. Kirtac and G. Germano, "Enhanced Financial Sentiment Analysis and Trading Strategy Development Using Large Language Models," Bangkok, Thailand, August 2024: Association for Computational Linguistics, in Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis https://doi.org/10.18653/v1/2024.wassa-1.1.
27. Liu X, et al. GPT understands too. AI Open. 2023. https://doi.org/10.1016/j.aiopen.2023.08.012.
28. Z. Du *et al.*, GLM: general language model pretraining with autoregressive blank infilling. 2022
29. J. Wei *et al.*, "Finetuned language models are zero-shot learners," *arXiv preprint* arXiv:2109.01652*,* 2021.
30. Gruver N, Finzi M, Qiu S, Wilson AG. Large language models are zero-shot time series forecasters. Adv Neural Inf Proc Syst. 2024;36:19622.
31. H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint* arXiv:2302.13971*,* 2023.
32. J. W. Rae *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint* arXiv:2112.11446*,* 2021.
33. Wei J, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37.
34. J. Liu *et al.*, "Generated knowledge prompting for commonsense reasoning," *arXiv preprint* arXiv:2110.08387*,* 2021.
35. Sattar MA, Toseef M, Sattar MF. Behavioral finance biases in investment decision making. Int J Account, Finance Risk Manag. 2020;5(2):69.

36. Chen W, El Majzoub A, Al-Qudah I, Rabhi FA. A CEP-driven framework for real-time news impact prediction on financial markets. Serv Orient Comput Applic. 2023. https://doi.org/10.1007/s11761-023-00358-8.
37. Mamaysky H. News and markets in the time of COVID-19. J Financ Quant Anal. 2023. https://doi.org/10.1017/S002210902300131X.
38. Chen W, Milosevic Z, Rabhi FA, Berry A. Real-time analytics: concepts, architectures, and ML/AI considerations. IEEE Access. 2023;11:71634–57. https://doi.org/10.1109/ACCESS.2023.3295694.
39. Li Y, Wang S, Ding H, Chen H. "Large language models in finance: a survey. Brooklyn: NY, USA; 2023.
40. A. Lopez-Lira and Y. Tang, "Can chatgpt forecast stock price movements? return predictability and large language models," *arXiv preprint* arXiv:2304.07619, 2023.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com